

**Comparing Multi-dimensional and Uni-dimensional
Computer Adaptive Strategies in Psychological and Health Assessment**

by

Jingyu Liu

BS, Beijing Institute of Technology, 1994

MS, University of Texas at San Antonio, 2002

Submitted to the Graduate Faculty of
School of Education in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

By

Jingyu Liu

It was defended on

June 4, 2007

and approved by

Susan Lane, Ph.D., School of Education

Feifei Ye, Ph.D., School of Education

James J. Irrgang, Ph.D., School of Health and Rehabilitation Sciences

Dissertation Advisor: Clement A. Stone, Ph.D., School of Education

Comparing Multi-dimensional and Uni-dimensional Computer Adaptive Strategies in Psychological and Health Assessment

Jingyu Liu, PhD

University of Pittsburgh, 2007

This study aimed to compare the efficiencies of multi-dimensional CAT versus uni-dimensional CAT based on the multi-dimensional graded response model and provide information about the optimal size of the item pool. Item selection and ability estimation methods based on multi-dimensional graded response models were developed and two studies, one based on simulated data, the other based on real data, were conducted. Five design factors were manipulated: correlation between dimensions, item pool size, test length, ability level, and number of estimated dimensions.

A modest effect due to the correlation between dimensions on the outcome measures was observed, although the effect was found primarily for correlations of 0 versus 0.4. Based on a comparison of the correlation condition equal to zero with correlation conditions greater than zero, the multi-dimensional CAT was more efficient than the uni-dimensional CAT. As expected, ability level had an impact on the outcome measures. A multi-dimensional CAT provided more accurate estimates for those examinees with average true ability values than those with true ability values in the extreme range. The multi-dimensional CAT was over-estimated for examinees with negative true ability values and under-estimated for examinees with positive true ability values. This is consistent with Bayesian estimation methods which “shrink” estimates toward the mean of the prior distribution. As the number of estimated dimensions increased, more accurate estimates were achieved. This supports the idea that the ability of one dimension can be used to augment the information available to estimate ability in another dimension. Finally, larger item pools and longer tests yielded more accurate and reliable ability estimation, although greater difference in efficiency was realized when comparing shorter tests and smaller item pools.

Information on the optimal item pool size was provided by plotting the outcome measures versus the item pool size. The plots indicated that, for short tests, the optimal item pool

size was 20 items; for longer test, the optimal item pool size was 50 items. However, if item exposure control or content balancing were an issue, a larger item pool would be needed to achieve the same efficiency in ability estimates.

TABLE OF CONTENTS

TPREFACE	XI
1.0 INTRODUCTION.....	1
2.0 REVIEW OF THE LITERATURE.....	5
2.1 ITEM RESPONSE THEORY	5
2.1.1 Uni-dimensional two-parameter model	6
2.1.2 Uni-dimensional graded response model	10
2.1.3 Multi-dimensional graded response model.....	15
2.2 COMPUTER ADAPTIVE TESTING	20
2.2.1 The CAT item pool.....	21
2.2.2 Starting the CAT	23
2.2.3 Stopping the CAT	24
2.2.4 Ability estimation and item selection methods.....	24
3.0 METHODOLOGY.....	30
3.1 ITEM SELECTION AND ABILITY ESTIMATION METHODS BASED ON MULTI-DIMENSIONAL GRADED RESPONSE MODEL.....	31
3.1.1 Ability estimation methods.....	31
3.1.1.1 MLE method	31
3.1.1.2 Bayesian-based method	33
3.1.2 Item selection methods	36
3.1.2.1 TMaximum information.....	37
3.1.2.2 Bayesian based method	39
3.2 STUDY 1.....	40
3.2.1 Experimental design	40
3.2.2 Outcome measures	45

3.3	STUDY 2.....	46
3.4	COMPUTER SIMULATION.....	47
3.4.1	Main procedures and subroutines.....	47
3.4.2	Validation.....	51
3.4.2.1	Validation of the generated item responses and ability estimation	51
3.4.2.2	Comparison with previous research	55
4.0	RESULTS	57
4.1	RESULTS FROM SIMULATED DATA	57
4.1.1	Root Mean Squared Error (RMSE) Measure	58
4.1.2	Bias Measure	70
4.1.3	Pearson Correlation Measure.....	79
4.1.4	Intra-class Correlation	84
4.1.5	Standard Error of estimate.....	88
4.1.6	Optimal size of the item pool.....	96
4.2	RESULTS FROM REAL DATA	102
4.2.1	Data Analyses	102
4.2.2	Outcome measures.....	105
5.0	DISCUSSION	108
5.1	SUMMARY OF FINDINGS.....	108
5.2	LIMITATIONS AND FUTURE RESEARCH.....	111
	APPENDIX A	114
	BIBLIOGRAPHY	135

LIST OF TABLES

Table 2.1. Sample CAT Item Pool.....	26
Table 2.2. Item Information Look-up Table	28
Table 2.3. Likelihood functions	28
Table 2.4. CAT results	29
Table 3.1. Item Parameters for Pool Size = 10	44
Table 3.2. One dimension: MCAT vs. MULTILOG	53
Table 3.3. Three dimensions: MCAT vs. NoHarm.....	54
Table 3.4. Correlations between nine dimensions	56
Table 3.5. Test Lengths.....	56
Table 3.6. Reliability.....	56
Table 3.7. Reliability from MCAT	56
Table 4.1. Root Mean Squared Error	59
Table 4.1. Continued.....	60
Table 4.2. Interpretation of n in Figure 1 ~ Figure 4	61
Table 4.3. ANOVA Results	70
Table 4.4. Bias	73
Table 4.4. Continued.....	74
Table 4.5. Pearson Correlation.....	80
Table 4.5. Continued.....	81
Table 4.6. Intra-class Correlation.....	86
Table 4.6. Continued.....	87
Table 4.7. Standard Error of Ability Estimates.....	88
Table 4.8. Standard Error of Estimates for all examinees	89

Table 4.9. Ability level based on dimension 1.....	90
Table 4.10. Ability level based on dimension 2.....	91
Table 4.11. Ability level based on dimension 3.....	92
Table 4.12. Interpretation of N in Figure 4.13 – Figure 4.15	93
Table 4.13. ANOVA Results	94
Table 4.14. Factor loadings for SF-36	103
Table 4.15. Item Pool for Real Data	104
Table 4.16. Root Mean Square Error	106
Table 4.17. Bias	107
Table 4.18. Pearson Correlation.....	107
Table 4.19. Intra-Class Correlation.....	107
Table 4.20. Standard Error of Estimates.....	107

LIST OF FIGURES

Figure 2.1. Item Characteristic Curve	6
Figure 2.2. ICC for different b parameters	7
Figure 2.3. ICC for different a parameters	8
Figure 2.4. ICCs and Information Curve	10
Figure 2.5. Category boundaries.....	11
Figure 2.6. Category Boundary Response Functions for a Five-Category Graded Item	12
Figure 2.7. Response Functions for a Five-Category Graded item	14
Figure 2.8a. Item Category Characteristic Surface	18
Figure 2.8b. Item Category Characteristic Surface	18
Figure 2.8c. Item Category Characteristic Surface	19
Figure 2.8d. Item Category Characteristic Surface	19
Figure 2.9. Iterative Process of CAT	21
Figure 2.10. Item Pool A.....	22
Figure 2.11. Item Pool B.....	23
Figure 3.1. Item Information Curves for Pool Size = 10	42
Figure 3.2. Flow Chart.....	50
Figure 3.3. Scatter Plots.....	52
Figure 4.1. Root Mean Squared Error under different correlation	62
Figure 4.2. Root Mean Squared Error under different test length	63
Figure 4.3. Root Mean Squared Error under different item pool size	64
Figure 4.4. Root Mean Squared Error under different ability level.....	655
Figure 4.5. Test Information for 10 items.....	68
Figure 4.6. Bias under different ability level	75

Figure 4.7. Bias under different test length.....	76
Figure 4.8. Bias under different item pool size.....	77
Figure 4.9. Bias under different correlation.....	78
Figure 4.10. Scatter plot for all ability groups.....	82
Figure 4.11. Scatter plot for examinees with true ability < -1	82
Figure 4.12. Scatter plot for examinees with $-1 < \text{true ability} < 0$	83
Figure 4.13. SE of estimates under different correlations between dimensions.....	95
Figure 4.14. SE of estimates under different test length.....	95
Figure 4.15. SE of estimates under different item pool sizes	95
Figure 4.16. Optimal Pool Size Curves for fixed-length tests	97
Figure 4.17. Optimal Pool Size Curves for variable-length tests	97
Figure 4.18. Optimal Pool Size for Standard Error of Estimates.....	99
Figure 4.19. Optimal Pool Size for Pearson Correlation	100
Figure 4.20. Optimal Pool Size for RMSE	101

PREFACE

A lot of people helped me reach this point and they deserve much gratitude for all of their assistance over the years.

First and foremost, I must thank my advisor Dr. Clement A. Stone for leading me into this wonderful program at the University of Pittsburgh and providing me with the unsurpassed guidance, training, and attention to detail during my time here. He is an excellent advisor and I am just so fortunate to be one of his students.

I am also grateful to my dissertation committee, Dr. Suzanne Lane, Dr. Feifei Ye, and Dr. James J. Irrgang, for their valuable input and suggestions, which enhanced the quality of this study.

I would also like to thank several family members. Without their help, the completion of this dissertation would not have been possible. Special thanks to my parents, Shumin Chen and Defu Liu, for their unconditional and unending support and love.

1.0 INTRODUCTION

In recent years, computerized tests have come to play an important role in the field of psychological assessment and health and medical sciences. For example, many popular personality tests are available in computer-administered versions, such as the California Psychological Inventory. Given the popularity of item response theory (IRT) and computer adaptive testing (CAT) in achievement tests, some researchers (Waller & Reise, 1989; McHorney, 1997) suggest that CAT should be used in personality assessment and generic health measurement. Waller and Reise (1989) also did a study on computerized adaptive personality assessment based on item response theory. Their results suggest that computerized adaptive personality assessment works very well. With the fixed-test-length strategy, a 50% savings in administered items was achieved with little loss of measurement precision.

However CAT has been ignored in these fields until very recently. One reason is that CAT is based on IRT, which more commonly assumes uni-dimensionality, but most psychological and health assessments are inherently multi-dimensional. These tests are often designed to provide comprehensive information along several dimensions of knowledge, attitude, or personality. As an example, the SF-36 survey is a general measure of health status that measures eight sub-domains of health. Research has demonstrated two general dimensions: physical and mental health dimensions (Haley, McHorney, & Ware, 1994; Ware, Kosinski, & Keller, 1994). Some tests, like GRE and SAT, are composed of several subtests that measure different abilities, and these abilities are usually not independent of one another. The ability of one dimension could help the examinees answer correctly on the items of another dimension. For example, an examinee exhibiting a high-level vocabulary proficiency is likely to exhibit a similar high-level of reading comprehension, and vice-versa. In this case, multi-dimensional IRT model should be used to model the item-examinee interaction.

In the late 1970s and early 1980s, a number of researchers were working actively to develop multi-dimensional IRT models. In addition to the work by Reckase (1972) on the multi-dimensional Rasch model, Mulaik (1972), Simpson (1978), and Whitely (1980) developed models for items with dichotomous responses. McKinley & Reckase (1982) considered many of the variations of the general Rasch model and decided that the linear logistic model was the most practical model. This model was labeled as a multivariate extension of the two-parameter logistic model. All of these studies concluded that multi-dimensional IRT models can estimate abilities in different dimensions simultaneously, and they also take into account the correlational structure among these abilities. Therefore, compared with uni-dimensional IRT models, multi-dimensional IRT models may be more accurate (Segall, 1996).

Segall (1996) presented maximum likelihood and Bayesian procedures for item selection and scoring of a multi-dimensional CAT based on dichotomous items. Segall compared a uni-dimensional CAT for nine power achievement subtests in the Armed Services Vocational Aptitude test battery to a multi-dimensional CAT. By maximizing the determinant of the posterior variance-covariance matrix as the statistical objective function for the multi-dimensional CAT item selection, Segall demonstrated detectable gains in the reliabilities of the outcome sub-scores when compared to simulated uni-dimensional CATs.

All of the above models are based on the dichotomous case. However, polytomous items are more commonly used in psychological and health measurement, such as the Likert-type scales traditionally used in questionnaires, attitude inventories, and surveys. In such items, a scale is used, and the labeling imparts the order, such as strongly disagree, disagree, indifferent, agree, and strongly agree. Under the graded scoring procedure, the item variable scale is divided into ordered categories. Samejima (1969, 1972) developed a graded response model to describe this kind of data. Dodd, Ayala, and Koch (1995) reviewed the research that has been conducted to investigate a variety of possible operational procedures for a polytomous model based CAT. They also conducted studies that compared polytomous CAT systems based on competing IRT models that are appropriate for the same measurement objective, as well as applications of polytomous CAT in marketing and educational psychology. Ayala (1994) discussed a multi-dimensional graded response model (MGRM) that is a direct generalization of Samejima (1969, 1972) uni-dimensional graded response model.

While the efficiencies of CAT in educational assessment with dichotomous items are well documented, the efficiencies associated with polytomous items are less well known (e.g., Stone & Irrgang, 2004; Ware, Bjorner, & Kosinski, 2000). This applies to both the cases of psychological and health assessment as well as to constructed response items in educational assessment. There is no research using CAT based on the multi-dimensional graded response model. One purpose of this paper is to compare the efficiencies of multi-dimensional CAT versus uni-dimensional CAT based on the graded response model for psychological and health assessment.

Another issue is the size of a CAT item pool. A CAT drawn from a large pool of items has a greater potential to meet the diverse needs of policy makers and researchers by administering fewer items but still providing precise estimation of the trait at all levels. A larger item pool can be achieved by including more items that measure a particular dimension or adding items that measure different correlated dimensions. The effect of item pool size on performance of a CAT with dichotomously scored educational tests has been studied. Stocking (1994) recommends that the CAT item pool for an exam program contain 12 times the number of items in an average CAT. For licensure and certification testing, Way (1998) suggests that a pool size of six to eight times the average CAT length might be adequate. However, the analogous impact in the context of psychological and health assessment has not received attention. Thus, another purpose of this study is to provide information about the optimal size of the item pool.

To achieve these goals, item selection and ability estimation methods based on multi-dimensional graded response models were developed in the present study, and Monte Carlo simulation studies were conducted. The specific research questions of this study included:

- 1) Is a multi-dimensional CAT more efficient than a uni-dimensional CAT?
- 2) How does the correlation between dimensions affect the efficiency of multi-dimensional CAT?
- 3) Is there any difference between the results at different levels of the trait?
- 4) What is the optimal size of the item pool?

The context of the present study is psychological and health assessment. However, it should be noted that the results may have implication for computer adaptive strategies with constructed response items in educational assessment. Future advances in computer scoring of

constructed response items may increase the use of computer adaptive testing with these types of items.

2.0 REVIEW OF THE LITERATURE

Item response theory (IRT) is a popular test theory and is currently an area of active research. IRT consists of a family of mathematical functions that model the interaction between a person responding to an instrument and items that are administered. The models predict performance on each item, or the probability of choosing a response category, using characteristics of items and persons. The person characteristics refer to the trait or traits (e.g., level of disability, ability, satisfaction) being measured, and the item characteristics refer to parameters used to describe the relationship between the item and the trait being measured.

2.1 ITEM RESPONSE THEORY

Item response theory (IRT) is a popular test theory and is currently an area of active research. IRT consists of a family of mathematical functions that model the interaction between a person responding to an instrument and items that are administered. The models predict performance on each item, or the probability of choosing a response category, using characteristics of items and persons. The person characteristics refer to the trait or traits (e.g., level of disability, ability, satisfaction) being measured, and the item characteristics refer to parameters used to describe the relationship between the item and the trait being measured.

A variety of IRT models have been developed for dichotomous and polytomous data. The most commonly used models for dichotomous items are the logistic models (e.g., two-parameter model and three-parameter model). Samejima's graded response model (Samejima, 1969, 1972) is applied to polytomous data, where options are ordered along continuum (e.g., likert scales). When there are only two response categories, Samejima's graded response model is identical to the two-parameter model. Therefore, the two-parameter model is a special case of Samejima's

graded response model. The early IRT applications involved primarily uni-dimensional IRT models. However, several multi-dimensional IRT models have been developed. These models usually are direct extensions of uni-dimensional models.

In this section, the two-parameter model is introduced first, and Samejima's model and multi-dimensional graded response model are then explained in detail and compared.

2.1.1 Uni-dimensional two-parameter model

The mathematical form of the two-parameter logistic model is given as:

$$P = \frac{1}{1 + \exp[-D * a(\theta - b)]} \quad (2.1)$$

where P is the probability of a correct response, and θ is the examinee's ability. The variables a and b are the item parameters. The item parameters vary from item to item, and they define the specific shape of the Item Characteristic Curve (ICC), which shows the probability of a correct response for students with different ability levels (Figure 2.1).

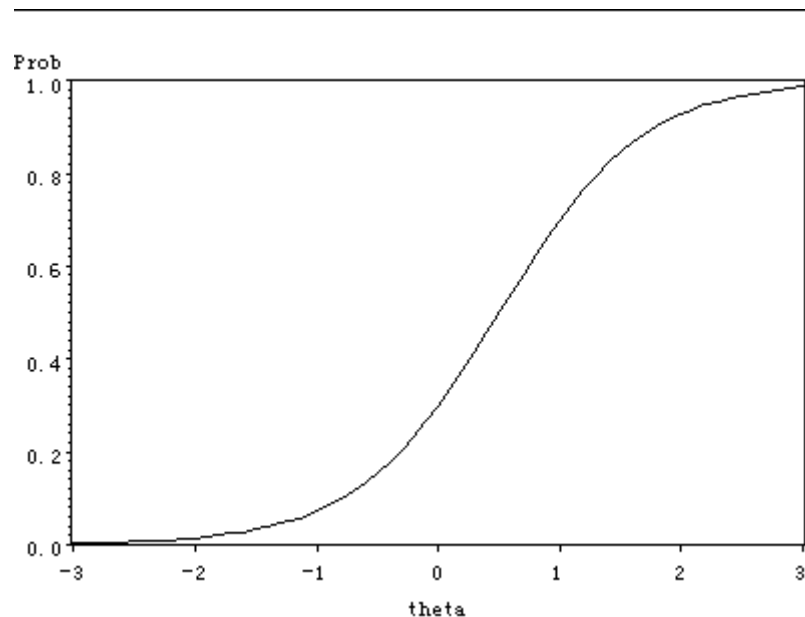


Figure 2.1. Item Characteristic Curve ($a = 1.0$, $b = 0.5$, $d = 1.70$)

The b parameter is called the difficulty parameter or the location parameter. The difficulty parameter sets the location of the curve on the horizontal axis; the curve shifts from left to right as the item becomes more difficult (Figure 2.2). The inflection point (point on the horizontal axis where the slope of the ICC is a maximum) on the curve will be at $P = .50$. When there is no guessing, b is the ability value where $P = .50$.

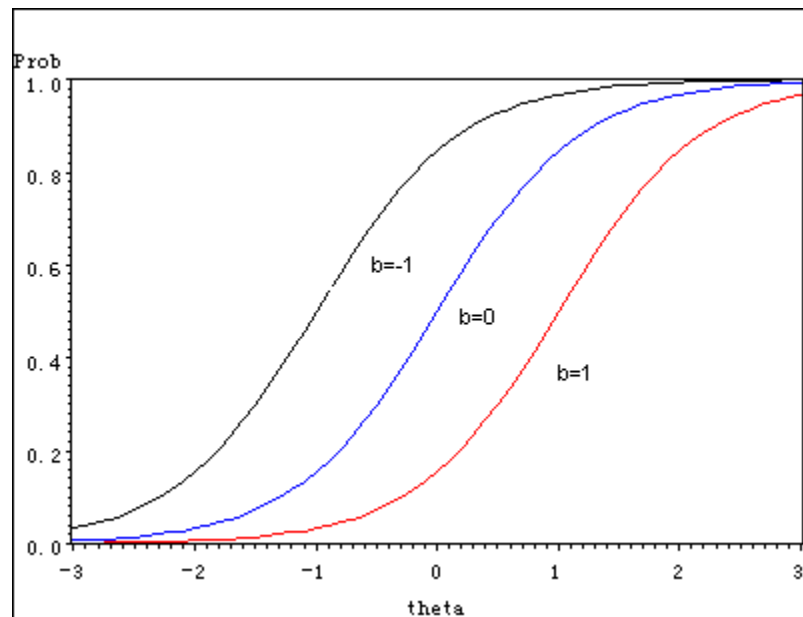


Figure 2.2. ICC for different b parameters ($a = 1.0$, $d = 1.70$)

The a parameter is called the discrimination parameter or the slope parameter. The a parameter is found by taking the slope of the line tangent to the ICC at b (Figure 2.3). The a parameter reflects the steepness of the curve at its steepest point. The steeper the curve, the more discriminating the item. As the a parameter decreases, the curve gets flatter until there is virtually no change in probability across the ability continuum. Items with very low a values are poor for distinguishing among examinees.

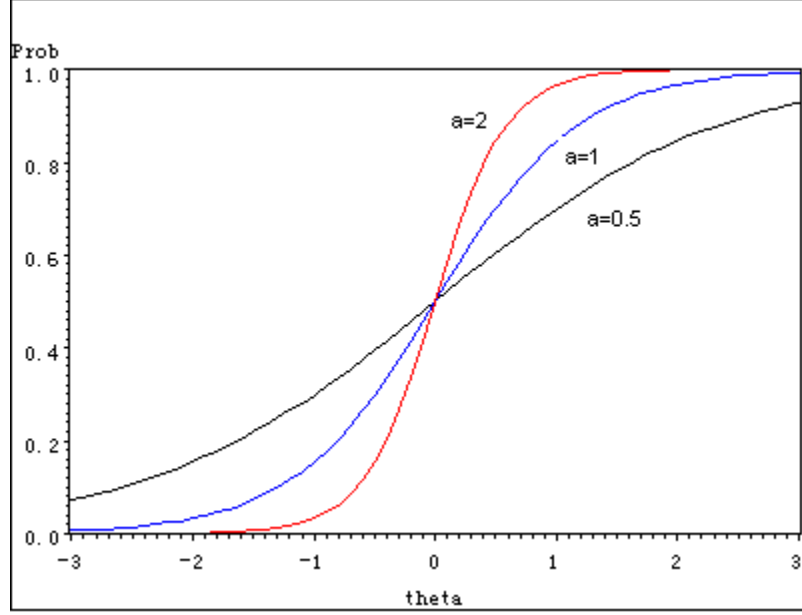


Figure 2.3. ICC for different a parameters ($b = 0.5, d = 1.70$)

One method for estimating the item and ability parameters involves maximizing the likelihood function of the observed item responses. The likelihood function for a given examinee of ability θ is the likelihood of a particular item response vector $U = (u_1, u_2, \dots, u_I)$, where $u_i = 1$ if the examinee answers the item i correctly, $u_i = 0$ otherwise. The likelihood function is given by

$$L(U|\theta) = \prod_{i=1}^I P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \quad (2.2)$$

where $P_i(\theta)$ is defined by Equation 2.1. For mathematical convenience, the logarithm of likelihood is used.

$$\log L(U|\theta) = \sum_{i=1}^I u_i \log[P_i(\theta)] + (1 - u_i) \log[Q_i(\theta)] \quad (2.3)$$

Another useful function in IRT is the item information function. When constructing a test form using IRT, item information functions are often used to select items. The item information is defined as

$$I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) \quad (2.4)$$

For two-parameter model, Equation 2.4 is simplified as

$$I(\theta) = a_i^2 P_i(\theta) Q_i(\theta) \quad (2.5)$$

The discrimination parameter plays an important role in the function. In general, highly discriminating items have tall, narrow information functions and they contribute greatly but over a narrow range. Less discriminating items provide less information but over a wider range. For example, Figure 2.4 shows the ICCs for two job satisfaction scale items that ask examinees to indicate whether their job is “satisfying” or “fascinating”. The top figure shows the two ICCs. The bottom figure shows the corresponding information functions. Both curves indicate that there are certain locations on the θ continuum where the items give maximum information and other locations where the items basically provide no information. Notice that the information function curve for “satisfying” is much higher, indicating that it gives more information at its maximum point. However, the item “fascinating” provides more information at the high, or very satisfied, end of the satisfaction continuum. This makes sense because only the most satisfied people will describe their job as fascinating, so the item “fascinating” will only discriminate among the most satisfied people. Similarly, an item, like “rotten”, would only discriminate among those who were very dissatisfied with their jobs. From Figure 2.4, we can see that the amount of item information depends upon the steepness of the ICC (the a parameter) and the location of the item information depends upon the difficulty or b parameter.

The test information function is simply the sum of the information functions of the items on the exam. Using this property with a large item bank, test information functions can be shaped to control measurement error very precisely.

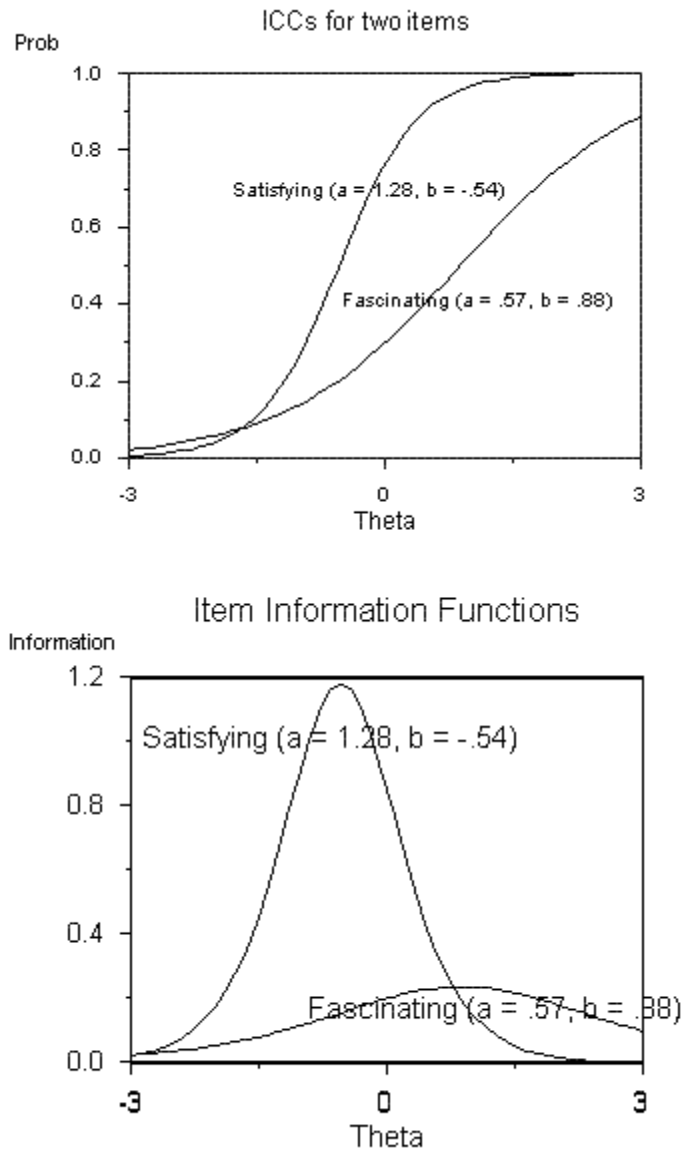


Figure 2.4. ICCs and Information Curve

2.1.2 Uni-dimensional graded response model

The graded response model (Samejima, 1969, 1972) is based on the category boundaries approach. Under the graded scoring procedure, the item variable scale is divided into ordered categories. Ordered categories are defined by boundaries that separate the categories. Logically,

there is always one less boundary than there are categories. Consider a specific item with five categories (labeled by strongly disagree, disagree, indifferent, agree, and strongly agree) separated by four category boundaries as in Figure 2.5. P^* is the probability that the response falls in or above the category.

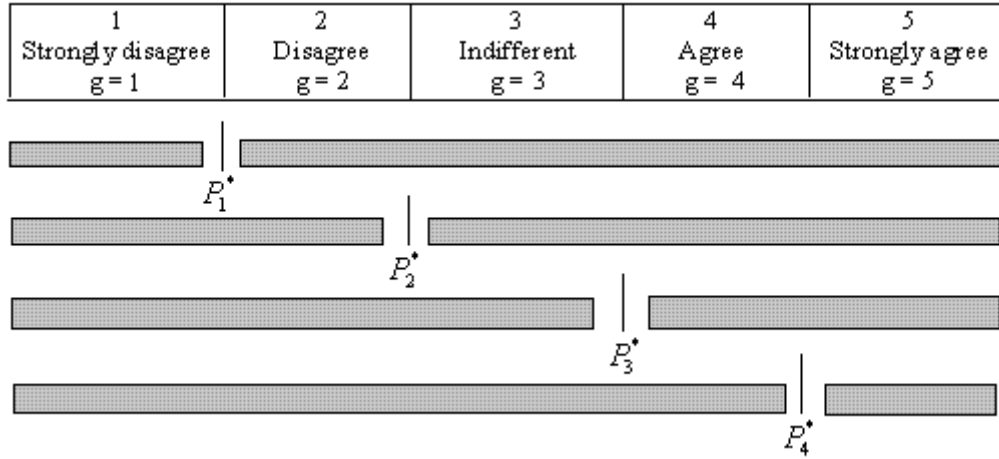


Figure 2.5. Category boundaries

The dichotomous item, which requires only one category boundary to separate the two possible response categories, is actually a special case for the graded response model. The key to making the graded response model work in practical terms is dichotomization. As shown in Figure 2.5, the first boundary between Category 1 and Category 2 divides all responses into two groups, those in or above this boundary (category 2, 3, 4, and 5), and those at or below this boundary (category 1). Similarly, the second, third, and fourth boundaries divide all responses into above or below the category boundary respectively. The Samejima graded response model (1969) is built on the two-parameter logistic (2PL) model because this dichotomous model is used as the function to obtain the cumulative probability which the response falls in or above the category g denoted by P_{ig}^* for the i th item. The usual equation for Samejima's graded response model is

$$P_{ig}^* = \frac{1}{1 + \exp[-Da_i(\theta - b_{ig})]} \quad (2.6)$$

where $D = 1.702$ is a constant, a_i is the item discrimination parameter, b_{ig} is the boundary location parameter, and θ is the examinee's ability. In order to maintain the underlying order of the response categories, the boundary location parameters must be ordered.

$$b_{m_i} > b_{m_i-1} > \dots > b_g > \dots > b_1 \quad (2.7)$$

For a five-category item, there will be four item characteristic curves or response functions for each boundary respectively as shown in Figure 2.6. For each curve, it is similar to the dichotomous case.

For Equation 2.6 to completely define an item's characteristic in terms of the available boundary functions, two additional definitions are required. The first is that the probability of responding in or above the lowest possible category for any item is defined as 1.0, across the entire range of θ . Algebraically, $P_{i0}^* = 1$. The probability of responding in a category higher than the highest category available, designated $P_{im}^*(\theta)$, is defined to equal zero throughout the trait range. Algebraically, $P_{im}^* = 0$.

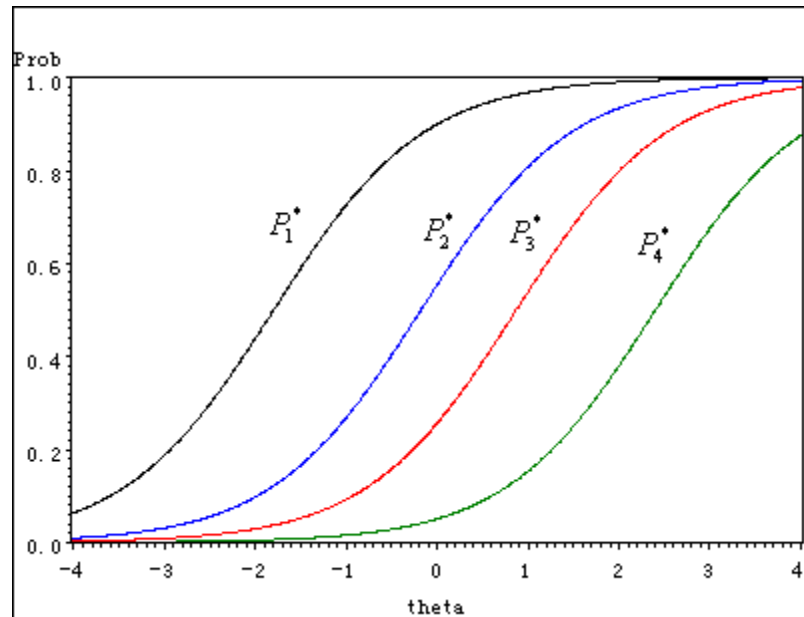


Figure 2.6. Category Boundary Response Functions for a Five-Category Graded Item
 $(a_i = 1.2209, b_1 = -1.8007, b_2 = -0.1872, b_3 = 0.8763, b_4 = 2.4070)$

We are more interested in the probability of responding in a given category, as this is the basis for determining respondents' levels of ability. If P_{ig} is the probability of responding in a particular category (g) to item i , it can be calculated as the difference between the cumulative probabilities of two adjacent categories:

$$P_{ig} = P_{ig-1}^* - P_{ig}^* \quad (2.8)$$

It is important to observe that, at each ability level,

$$\sum_{g=1}^{m_i} P_{ig}(\theta) = 1 \quad (2.9)$$

Figure 2.7 depicts the probability functions of responding in a given category for a graded item having five categories with $a_i = 1.2209$, $b_1 = -1.8007$, $b_2 = -0.1872$, $b_3 = 0.8763$, $b_4 = 2.4070$. It is apparent that the category response functions are no longer monotonic functions. Only the functions for the extreme categories are monotonically decreasing and increasing respectively. The curves for the center three response categories, $g = 2, 3, 4$, are unimodal functions of θ . These curves will be referred to as Item Category Characteristic Curves (ICCC).

The likelihood function for a given examinee of ability θ is the likelihood of a particular item response vector $U = (u_{1g}, u_{2g}, \dots, u_{ng})$, where $u_{ig} = 1$ if the examinee chose response g to item i , $u_{ig} = 0$ otherwise. The likelihood function is given by

$$L(U|\theta) = \prod_{i=1}^N \prod_{g=1}^{m_i} P_{ig}^{u_{ig}} \quad (2.10)$$

where $P_{ig}(\theta)$ is defined by Equation 2.6. For mathematical convenience, the logarithm of likelihood will be used.

$$\ln L(U | \theta) = \sum_{i=1}^N \sum_{g=1}^{m_i} u_{ig} \log P_{ig} \quad (2.11)$$

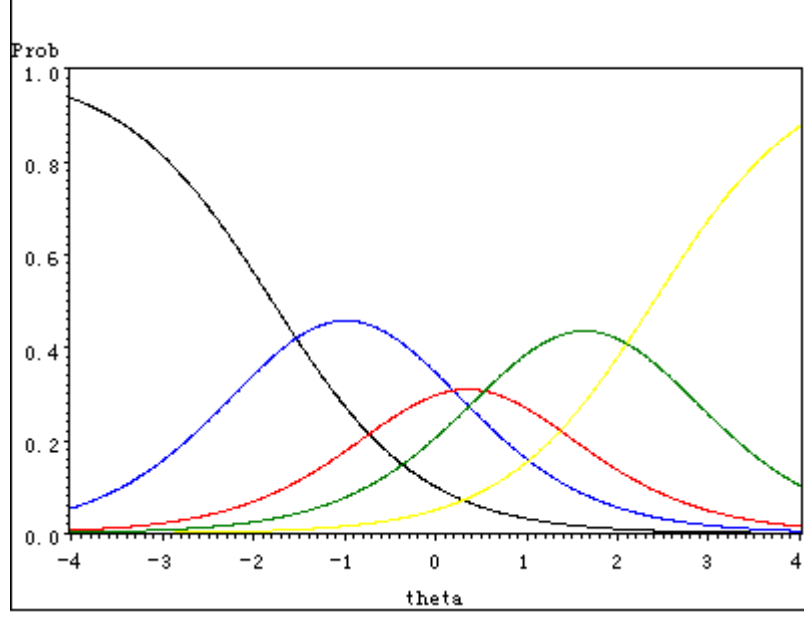


Figure 2.7. Response Functions for a Five-Category Graded item
 $(a_i = 1.2209, b_1 = -1.8007, b_2 = -0.1872, b_3 = 0.8763, b_4 = 2.4070)$

Samejima (1969, 1972) has defined the information of an item response category as

$$I_{ig}(\theta) = -\frac{\partial^2 \log P_g(\theta)}{\partial \theta^2}, \quad (2.12)$$

which is equivalent to

$$I_{ig}(\theta) = \left\{ \frac{P'_{ig}(\theta)}{P_{ig}(\theta)} \right\}^2 - \frac{P''_{ig}(\theta)}{P_{ig}(\theta)} \quad (2.13)$$

Because an item's categories are not independent of one another, it is not possible to simply define item information as a sum of category information. Instead, an intermediary term must be constructed that represents the share of information that each category provides to item information. Category share information is obtained as a weighted component of category information, where I_{ig} is weighted by the response probability for a category across θ . Category share information is therefore designated

$$I_{ig}(\theta)P_{ig}(\theta) \quad (2.14)$$

Item information is then a simple sum of category share information across all the categories of an item.

$$I_i = \sum_{g=1}^{m_i} I_{ig} P_{ig} \quad (2.15)$$

Substituting 2.13 into 2.15,

$$I_i = \sum_{g=1}^{m_i} \left[\frac{(P'_{ig})^2}{P_{ig}} - P''_{ig} \right] = \sum_{g=1}^{m_i} \left[\frac{(P_{ig-1}^{*'} - P_{ig}^{*'})^2}{P_{ig-1}^* - P_{ig}^*} - (P_{ig-1}^{*''} - P_{ig}^{*''}) \right] \quad (2.16)$$

Samejima (1969) showed that the sum of the second derivatives of the Equation 2.10 for an item equal zero. Therefore, Equation 2.16 can be simplified to

$$I_i = \sum_{g=1}^{m_i} \frac{(P_{ig-1}^{*'} - P_{ig}^{*'})^2}{P_{ig-1}^* - P_{ig}^*} \quad (2.17)$$

2.1.3 Multi-dimensional graded response model

Ayala (1994) discussed a multi-dimensional graded response model (MGRM), which is a direct generalization of the uni-dimensional graded response model. The difference is that the parameters are single scalar values for the uni-dimensional model, but vectors for the multi-dimensional model. In the MGRM, examinee responses to item i are categorized into m_i ordered categories in which higher categories indicate greater θ level and m_i is the number of category boundaries. The MGRM is expressed as

$$\begin{aligned} P_{ig}^*(\Theta) &= \frac{1}{1 + \exp[-DA_i'(\Theta - B_{ig})]} \\ &= \frac{1}{1 + \exp[-D \sum_{k=1}^h a_{ik}(\theta_k - b_{ig})]} \end{aligned} \quad (2.18)$$

where

D is a scaling constant,

θ_k is the latent trait on dimension k ($k = 1, \dots, h$ dimensions),

a_{ik} is the discrimination parameter for item i on dimension k ,

b_{ig} is the difficulty parameter for category g for item i , and the summation is across all dimensions.

In the case of h dimensions, Θ is an $h \times 1$ vector. A_i is an $h \times 1$ vector. B_{ig} is an $h \times 1$ vector.

$$\Theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_h \end{bmatrix}, \quad A_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ih} \end{bmatrix}, \quad B_{ig} = \begin{bmatrix} b_{ig} \\ b_{ig} \\ \vdots \\ b_{ig} \end{bmatrix}$$

$P_{ig}^*(\Theta)$ is the probability of a randomly selected examinee with latent traits Θ responding in category g or higher for item i . As for the uni-dimensional model, $P_0^* = 1$ and $P_{m_i}^* = 0$, and the probability of responding in a particular category g can be calculated as the difference between the cumulative probabilities of two adjacent categories:

$$P_{ig} = P_{ig-1}^* - P_{ig}^* \quad (2.19)$$

When $h \geq 2$ and $m_i = 2$, the MGRM reduces to the multi-dimensional two-parameter logistic model (Mckinley & Rechase, 1983). If $h = 1$, the MGRM reduces to the graded response model. When $h = 1$ and $m_i = 2$ (correct and incorrect), the MGRM model reduces to the two-parameter model.

There are three types of vectors with parameters in Equation 2.18. They are the ability parameter vector Θ , and item parameter vectors, A and B_g . The ability parameter vector, Θ , and the difficulty parameter vector, B_g , have the same meanings as for the uni-dimensional IRT model. But the parameter vector A doesn't have exactly the same meaning as with the uni-dimensional model.

Elements of A are related to the discriminating power of each test item. The A -vector can be interpreted in a similar way as the a parameter in uni-dimensional IRT models. The elements of the vector are related to the slope of the item response surface in the direction of the corresponding θ axis. The elements therefore indicate the sensitivity of the item to differences in ability along the θ axis. However, the discriminating power of an item differs depending on the direction that is being measured in the Θ space. If the direction of interest in the space is parallel to the surface, the slope will be zero, and the item is not discriminating. Unless an item is a pure measure of a particular dimension, it will be more discriminating for combinations of dimensions

than for single dimensions. The discriminating power of the item for the most discriminating combinations of dimensions is given by

$$MDISC_i = \text{sqr}t\left(\sum_{k=1}^h a_{ik}^2\right) \quad (2.20)$$

where $MDISC_i$ is the discrimination of the item i for the best combination of abilities; h is the number of dimensions in the Θ space; and a_{ik} is an element of the A vector.

The uni-dimensional graded response model uses Item Category Characteristic Curves (ICCC) to describe graphically the functional relationship between the probability of a correct response and the underlying ability. The ICCC generalizes to an item category characteristic surface (ICCS) for the multi-dimensional IRT model. When there are only two dimensions, the form of the probability surface can be represented graphically as Figure 2.8a-2.8e. Each curve in Figure 2.7 becomes a surface in Figure 2.8a-2.8e. The surfaces in Figure 2.8a and Figure 2.8e correspond to the curves of P_1 and P_5 in Figure 2.7. They are monotonically increasing or decreasing when both θ_1 and θ_2 become smaller. The surfaces in Figure 2.8b, 2.8c, and 2.8d correspond to the center three curves of P_2 , P_3 and P_4 in Figure 2.7. They are unimodal functions of θ_1 and θ_2 .

The likelihood function has the same form as the uni-dimensional model (Equation 2.10). The item information function is an $h \times h$ information matrix, denoted by $I(\Theta)$.

$$I(\Theta) = \begin{bmatrix} \sum_{g=1}^{m_i} \frac{(\partial P_{ig} / \partial \theta_1)^2}{P_{ig}} & \dots & \sum_{g=1}^{m_i} \frac{\partial P_{ig} / \partial \theta_1 \times \partial P_{ig} / \partial \theta_h}{P_{ig}} \\ \vdots & \ddots & \vdots \\ \sum_{g=1}^{m_i} \frac{(\partial P_{ig} / \partial \theta_h)^2}{P_{ig}} \end{bmatrix} \quad (2.21)$$

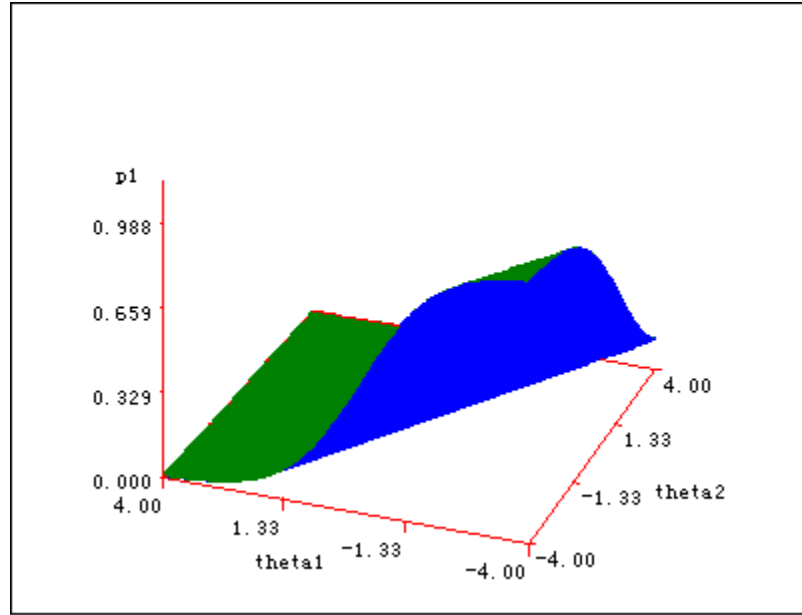


Figure 2.8a. Item Category Characteristic Surface (P_1)

$$a_{i1} = 0.5, a_{i2} = 1.0, a_{i3} = 2.0, b_1 = -1.8007, b_2 = -0.1872, b_3 = 0.8763, b_4 = 2.4070$$

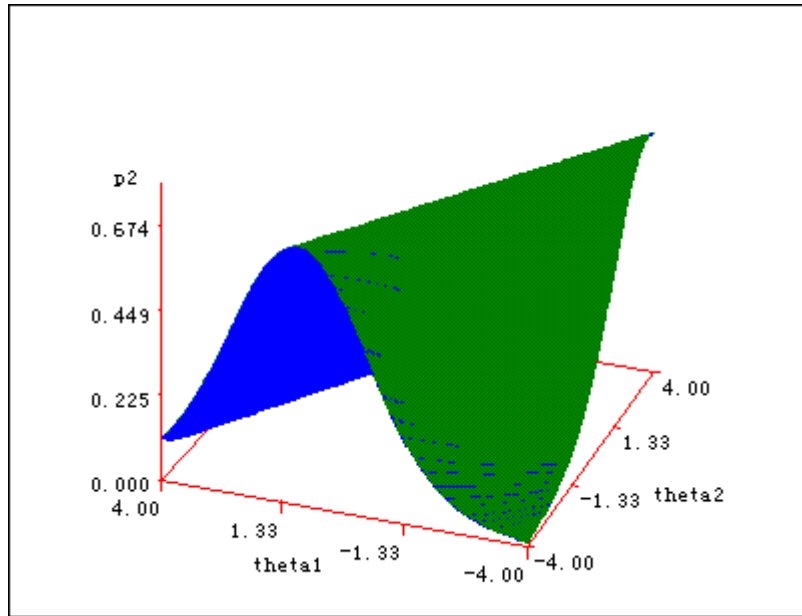


Figure 2.8b. Item Category Characteristic Surface (P_2)

$$a_{i1} = 0.5, a_{i2} = 1.0, a_{i3} = 2.0, b_1 = -1.8007, b_2 = -0.1872, b_3 = 0.8763, b_4 = 2.4070$$

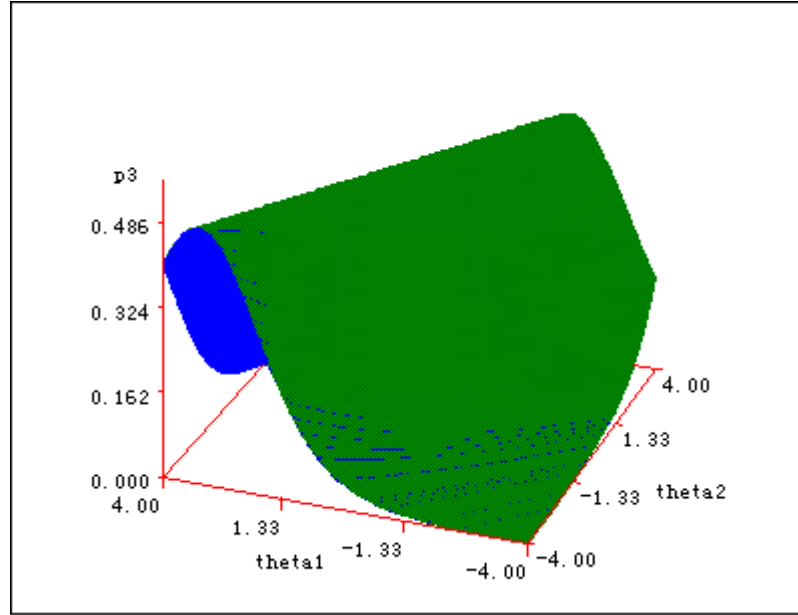


Figure 2.8c. Item Category Characteristic Surface (P₃)

$$a_{i1} = 0.5, a_{i2} = 1.0, a_{i3} = 2.0, b_1 = -1.8007, b_2 = -0.1872, b_3 = 0.8763, b_4 = 2.4070$$

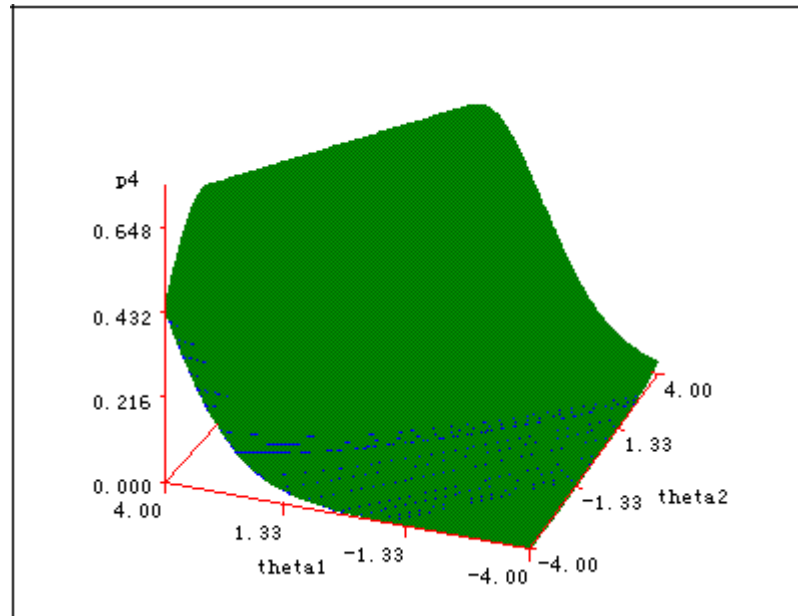


Figure 2.8d. Item Category Characteristic Surface (P₄)

$$a_{i1} = 0.5, a_{i2} = 1.0, a_{i3} = 2.0, b_1 = -1.8007, b_2 = -0.1872, b_3 = 0.8763, b_4 = 2.4070$$

2.2 COMPUTER ADAPTIVE TESTING

CAT has been popular for the past two decades. With computer adaptive tests, the computer selects and presents test items to examinees according to the estimated level of the examinee's ability. Examinees can be given the items that maximize the information about their ability levels from their item responses. Thus, examinees will typically receive fewer items that are very easy or very hard. This tailored item selection results in reduced standard errors and greater precision with properly selected items. CAT exams usually contain fewer items than conventional paper-and-pencil measures. On average, a CAT exam is 50% shorter than a paper-and-pencil measure with equal or better measurement precision (Wainer, 2000).

Most CAT exams are based on an item pool or bank that has been scaled using item response models. Either using uni-dimensional CAT or multi-dimensional CAT, the iterative process has the following steps:

1. All the items in the bank that have not yet been administered are evaluated to determine which will be the best one to administer next given the current ability level estimate.
2. The item that provides the maximum information is administered and the examinee responds.
3. A new ability estimate is computed based on the responses to all of the administered items.
4. Steps 1 through 3 are repeated until a stopping criterion is met.

The iterative process can be described by Figure 2.9.

Two important steps in CAT algorithms are item selection and ability estimation. Wainer (1990) has demonstrated that we learn little about an examinee's ability if we persist in asking questions that are far too difficult or far too easy for that person. We learn the most about an examinee's ability when we accurately direct our questions at the current level of the examinee's ability. The item selection algorithm needs to make sure the most informative test items are selected from an item pool, so that each examinee's ability can be efficiently estimated with a short test.

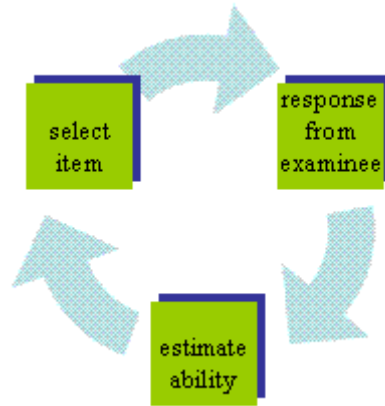


Figure 2.9. Iterative Process of CAT

Since an IRT scaled item pool is used, the only parameter that has to be estimated during the computerized test is the examinee's latent ability, θ . Ability estimates are updated following each item response to provide the current best estimate of an examinee's true ability.

In addition to item selection and ability estimation methods, there are other aspects that need to be considered. These include the pool or bank of items for the CAT, selecting the first item, and stopping the CAT. These issues are discussed next.

2.2.1 The CAT item pool

An item pool is a collection of test items that can be used to assemble or construct a CAT. The item pool should include sufficient numbers of items of satisfactory quality, and appropriately targeted to examinee ability. It is beneficial to review the overall quality of the item pool in terms of maximum item information, because most items are selected for administration on the basis of an item's estimated maximum information.

Requirements for the item pool are related to the test purpose, the test delivery method, the measurement or statistical assumptions of the model used to obtain the item characteristics within the pool, the test length, the frequency of test administrations, and security requirements. For example, a norm-referenced test is typically designed to maximize the range over which ability scores are obtained. The desired shape of the target test information function should be

centered on the mean and provide information for abilities from -2.0 to 2.0. In contrast, a pass/fail test would have a peak at a cut-score where passing is defined.

Examples of test information plots appear in Figure 2.10 and 2.11. The maximum information plot for Pool A illustrates a pool of items that are dispersed fairly well and possess items that have relatively high maximum information across a broad range of ability. This pool could be suited for a CAT, in which the assessment goal is to estimate an examinee's ability as accurately as possible.

The plot for Pool B is different from that of Pool A in the distributional characteristics of maximum item information. Pool B contains items with smaller values of maximum information; they also are concentrated in the area of θ from -2.0 to -1.0. Based on this plot, Pool B may be better suited for a classification test with a cut score between -2.0 and -1.0. There is a higher density of items around any latent passing score in the interval [-2.0, -1.0] than other areas on the ability metric.

An effective CAT should quickly identify and administer items that are most informative at an estimate of the examinee's true ability. Usually the distribution of the examinee's true ability follows a normal distribution. Therefore, for a high quality CAT item pool, items should be evenly and equally distributed throughout the θ continuum of interest (Urry, 1977; Weiss, 1982).

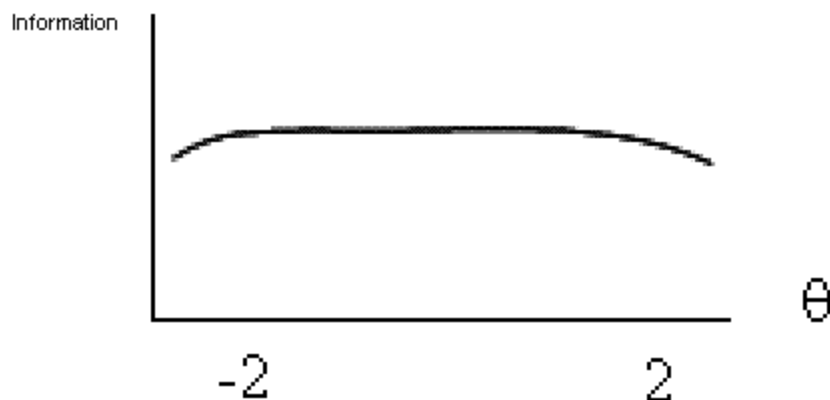


Figure 2.10. Item Pool A



Figure 2.11. Item Pool B

2.2.2 Starting the CAT

The starting point for a CAT refers to the difficulty level of the initial item or items administered to an examinee. Parshall et. al. (2002) summarized three approaches that may be used to select these initial items.

The “best guess” approach administers an item of medium difficulty on the grounds that if we know nothing about an examinee, our best guess is that he or she is like most other examinees. If the examinee population is normally distributed, a reasonable choice for starting a CAT is with an item of moderate difficulty, such as one with a difficulty parameter between -0.5 and 0.5 .

The “start easy” method begins the test with relatively easy items in order to give the examinee time to “warm up.”

Finally, in the “use what you’ve got” method, other test scores or information can be used to refine the initial estimate of examinee ability and thus the most appropriate level of item difficulty. This is also called “Collateral Data” method. In this method, an empirical predictor of the θ s of the examinees derived from information on available background variables when the test is administered. For example, examinees frequently provide biographical information when registering for a CAT session, which can be used to derive a statistical relation between such data

and the θ s of the examinees. Also, most CAT sessions begin with the examinees reading the instructions and responding to a few sample exercises. The time used to work through the instructions and/or the exercises may contain statistical information on the examinee's ability. Another example would be scores on previous attempts to pass the same test. These previous scores can be used to predict examinees' θ .

2.2.3 Stopping the CAT

Conventional paper-pencil tests are fixed-length tests, in which all examinees have the same number of items. A fixed-length test can also be used to depict a stopping rule for CAT tests. In addition, a CAT can have a variable-length test.

For fixed-length tests, different examinees may be tested at different levels of precision. Examinees who are more easily "targeted" by their selected test, either because they respond more predictably or because their ability falls where the CAT item pool is strong, are measured more precisely than poorly targeted examinees.

In contrast, a variable-length CAT stops when an examinee gets to a fixed level of precision, which results in administering different numbers of items to different examinees. Well-targeted examinees generally receive shorter tests than poorly targeted examinees. Variable-length testing can result in biased estimates of ability, especially if the test is short (Parshall et. al., 2002). In addition, a variable-length stopping rule would be hard to explain to a lay audience. Therefore, a fixed-length CAT is more popular.

2.2.4 Ability estimation and item selection methods

After an item is administered, there is more information about a person's standing on the trait of interest, and a point estimate of the person's standing on the trait and a confidence interval about the estimate can be attained. At the end of the test, a final estimate of the person's standing and a final confidence interval around that point can be attained. In CAT, the standard error of measurement can be used to evaluate the error in estimating the current ability estimate.

After the information for the set of items administered to the candidate is calculated, the standard error of measurement for examinee j 's estimated ability level is then defined by

$$SE = \frac{1}{\sqrt{I_j}} \quad (2.22)$$

where I_j is the test information for the specific test administered to examinee j .

Three popular ability estimation methods are maximum likelihood estimation (MLE), maximum a posteriori (MAP), and expected a posteriori (EAP). These last two are Bayesian-based methods and can be referred to as the Bayes mode and Bayes mean approaches, respectively. The maximum likelihood estimate of ability is determined by finding the mode or maximum value of the likelihood function (Equation 2.3 for dichotomous model, Equation 2.11 for graded response model). Bayesian methods apply Bayes theorem. For the CAT application, it is defined by

$$P(\theta|U) \propto L(\theta|U)P(\theta), \quad (2.23)$$

where the probability $P(\theta)$ is called the prior distribution of latent ability. The prior distribution expresses what is known about how the latent ability is distributed in the tested population before the test is administered. $L(\theta|U)$ is the likelihood function. $P(\theta|U)$ is the posterior distribution. When estimating ability, the posterior distribution is used instead of the likelihood function. In the EAP, or Bayes mean approach, the mean of the posterior distribution is computed as the point estimate of ability. In the MAP, or Bayes mode approach, the mode or maximum value taken on by the posterior is used.

The goal of CAT item selection is to maximize efficiency and produce a short, informative test for each examinee. Generally, there are two item selection methods: maximum information and Bayesian based method. Under the maximum information item selection procedure, the item that has the largest information value at the examinee's current ability estimate is selected for administration. The item information function is defined by Equation 2.5 for the dichotomous model, Equation 2.17 for the uni-dimensional graded response model, and Equation 2.21 for the multi-dimensional graded response model. Under the Bayesian based method, the general weighted information criterion (GWIC) is used. The GWIC uses information values over a posterior θ distribution, instead of a single value at a specific θ level. These values

are aggregated into a single value using a weighted average (Equation 2.24). The item having the maximum weighted average is selected for the examinee:

$$I(\theta | U) = \int_{-\infty}^{\infty} I(\theta)P(\theta | U)d\theta \quad (2.24)$$

A simple example can illustrate the philosophy of the CAT method. In this example, maximum information and maximum likelihood function are used as the item selection and ability estimation methods, respectively. Consider a 10-item pool with the item parameter estimates from a two-parameter model given in Table 2.1. Before the test is administered to any examinee, an information table (Table 2.2) is constructed based on parameter estimates for items in the item pool. For each item in the pool, the values of θ are listed in increments of 0.5 across the entire ability range from -4.0 to 4.0. During a CAT, whenever a new item must be selected for administration, the table is used in a look-up fashion to find the item in the pool that has maximum information at a value of θ that is closest to the current estimate of θ . Basically, this is how most CAT algorithms locate, select, and administer items in a CAT. The major idea behind the creation of such a table is to make the increments small enough so that they represent the continuous information function, $I(\theta)$, but large enough so that the table is still manageable.

Table 2.1. Sample CAT Item Pool

Item	a	b
1	0.397	-2.237
2	0.537	-1.116
3	1.440	1.496
4	0.920	0.801
5	1.261	-0.469
6	0.857	-0.103
7	1.471	0.067
8	1.382	0.495
9	0.940	0.801
10	1.290	1.170

The steps taken for the administration of this sample CAT are as follows:

Step 1. The “best guess” method is used to start the test. Based on Table 2.2, item 7 should be administered to this examinee because this item has the largest amount of information for examinee with $\hat{\theta} = 0.0$.

Step 2. The examinee provides a correct answer to item 7. So $U = (1)$.

Step 3. $L(U | \theta) = P_1(\theta)$. The likelihood function (LF1) for the different values of θ for item 7 is given in Table 2.3. The value of θ that gives the maximum likelihood of a single correct response, in terms of the θ values given in the table, is $\hat{\theta} = 4.0$.

Step 4. Based on Table 2.2, the next item that should be administered to this examinee is item 4 because this item has the largest amount of information for examinees with $\hat{\theta} = 4.0$.

Step 5. This time, the examinee provides an incorrect answer to item 4. So $U = (1, 0)$.

Step 6. Based on the updated likelihood function (LF2), which is $L(U | \theta) = P_1(\theta)Q_2(\theta)$, the maximum of this likelihood occurs for an ability estimate of $\hat{\theta} = 0.5$.

⋮

Last step. Assume that the stopping rule is fixed-length test, and the length is 4. The test is stopped when four items have been administered. The results of this simple example are summarized in Table 2.4.

The basic idea of selecting items and estimating ability for a multi-dimensional CAT is similar to a uni-dimensional CAT. However, the problems of CAT item selection and ability estimation become more complex in a multi-dimensional context. For example, unlike a uni-dimensional CAT, which merely administers items targeted to an examinee’s location along a single score scale, a multi-dimensional CAT must locate an examinee’s ability estimates on a plane or hyper-plane and administer items that ideally minimize the joint estimation errors for those estimates. The item selection and ability estimation methods based on multi-dimensional graded response model are discussed in next section.

Table 2.2. Item Information Look-up Table

θ	Item									
	1	2	3	4	5	6	7	8	9	10
-4	0.179	0.018	0.000	0.000	0.001	0.002	0.000	0.000	0.000	0.000
-3.5	0.033	0.026	0.000	0.001	0.002	0.005	0.000	0.000	0.001	0.000
-3	0.037	0.037	0.000	0.002	0.007	0.010	0.001	0.001	0.002	0.000
-2.5	0.039	0.050	0.000	0.005	0.020	0.021	0.003	0.002	0.004	0.001
-2	0.039	0.061	0.000	0.010	0.055	0.041	0.012	0.005	0.010	0.002
-1.5	0.037	0.070	0.001	0.022	0.141	0.075	0.041	0.017	0.021	0.005
-1	0.033	0.072	0.005	0.045	0.292	0.123	0.131	0.054	0.044	0.014
-0.5	0.028	0.067	0.015	0.086	0.397	0.169	0.339	0.153	0.087	0.040
0	0.023	0.056	0.050	0.146	0.312	0.183	0.537	0.346	0.150	0.110
0.5	0.019	0.044	0.153	0.200	0.157	0.152	0.409	0.477	0.209	0.253
1	0.014	0.032	0.366	0.207	0.062	0.102	0.174	0.342	0.215	0.402
1.5	0.011	0.022	0.518	0.159	0.023	0.059	0.057	0.150	0.164	0.366
2	0.008	0.015	0.362	0.097	0.008	0.031	0.017	0.052	0.099	0.199
2.5	0.006	0.010	0.150	0.052	0.003	0.016	0.005	0.017	0.051	0.081
3	0.004	0.006	0.049	0.025	0.001	0.008	0.001	0.005	0.025	0.029
3.5	0.003	0.004	0.015	0.012	0.000	0.004	0.000	0.002	0.011	0.010
4	0.002	0.003	0.004	0.006	0.000	0.002	0.000	0.001	0.005	0.003

Table 2.3. Likelihood functions

θ	Likelihood Function			
	LF1	LF2	LF3	LF4
-4	0.000	0.000	0.000	0.000
-3.5	0.000	0.000	0.000	0.000
-3	0.000	0.000	0.000	0.000
-2.5	0.002	0.002	0.000	0.000
-2	0.006	0.006	0.000	0.000
-1.5	0.019	0.019	0.000	0.000
-1	0.065	0.061	0.002	0.000
-0.5	0.195	0.172	0.015	0.000
0	0.458	0.356	0.085	0.002
0.5	0.747	0.460	0.231	0.019
1	0.912	0.385	0.295	0.068
1.5	0.973	0.244	0.223	0.112
2	0.992	0.132	0.128	0.099
2.5	0.998	0.065	0.065	0.060
3	0.999	0.031	0.031	0.030
3.5	1.000	0.014	0.014	0.014
4	1.000	0.007	0.007	0.007

Table 2.4. CAT results

Order	Item	Item Response	$\hat{\theta}$
1	7	1	4.0
2	4	0	0.5
3	8	1	1.0
4	3	1	1.5

3.0 METHODOLOGY

Although the efficiencies of CAT and the effect of item pool size on performance of a CAT with dichotomously scored tests in educational assessment have been studied and are well documented, there is no research about CAT based on the multi-dimensional graded response model. The purposes of this study are three fold: 1) extend item selection and ability estimation methods to the multi-dimensional graded response model; 2) compare the efficiencies of multi-dimensional CAT versus uni-dimensional CAT based on the graded response model; and 3) provide information about the optimal size of the item pool for multi-dimensional CAT. Analyses based on a Monte Carlo study and real data were employed to achieve these purposes.

The context of the present study is psychological and health assessment. However, it should be noted that, the results may have implication for computer adaptive strategies with constructed response item in educational assessment. Future advances in computer scoring of constructed response items may evaluate the use of computer adaptive testing with these types of items.

Two studies were designed to compare uni-dimensional and multi-dimensional CAT strategies and explore features of the design of the CAT that may affect test performance. One study uses simulated data and the other study uses real data. For both studies, a program was developed to simulate computer adaptive testing in the context of psychological and health assessments. In this type of research, it is typically the case that several assessments, each measuring more than one construct, are administered to subjects. Thus, a multi-dimensional structure to the different assessments was assumed. Further, a simple structure, where each item loads on a single dimension, was assumed. Consistent with the type of items in these assessments, the IRT model used was a multi-dimensional graded response model.

The algorithm of the CAT computer program for these studies follows the process described in Chapter 2. Only Bayesian based methods were used since these methods considered

the correlations between different dimensions. For these studies, only fixed-length CATs were considered since these are currently used in most CAT assessments. As for starting the CAT, all examinees were assumed to be “average”, and an item of moderate difficulty was selected as the first item. Other issues, such as exposure control, are not problems for psychological and health assessment, and they were not considered in this design.

3.1 ITEM SELECTION AND ABILITY ESTIMATION METHODS BASED ON MULTI-DIMENSIONAL GRADED RESPONSE MODEL

3.1.1 Ability estimation methods

In a multi-dimensional CAT, the only parameters that have to be estimated are the examinee’s latent abilities, $\{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_h\}$. Ability estimates are updated following each item response to provide the current best estimate of an examinee’s true abilities. The popular ability estimation methods are maximum likelihood estimation (MLE) and Bayesian-based procedures. The multi-dimensional ability estimation methods are simply extensions from uni-dimensional methods.

3.1.1.1 MLE method

The vector of values $\{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_h\}$ that maximize the likelihood function given by Equation 2.11 is taken as the estimator of the vector Θ . The ML estimates are the solutions to the set of h simultaneous equations given by

$$\frac{\partial}{\partial \Theta} \log L(U | \Theta) = 0 \quad (3.1)$$

where

$$\frac{\partial}{\partial \Theta} \log L(U | \Theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \log L(U | \Theta) \\ \frac{\partial}{\partial \theta_2} \log L(U | \Theta) \\ \vdots \\ \frac{\partial}{\partial \theta_h} \log L(U | \Theta) \end{bmatrix} \quad (3.2)$$

Explicit expression for these partial derivatives can be obtained by first noting that the natural logarithm of the likelihood function (Equation 2.11) is

$$\log L(U|\Theta) = \sum_{i=1}^n \sum_{g=1}^{m_i} u_{ig} \log P_{ig} \quad (3.3)$$

The derivative of the log likelihood with respect to θ_k (*for* $k = 1, 2, \dots, h$) takes on a form similar to the uni-dimensional graded response model:

$$\frac{\partial}{\partial \theta_k} \log L(U | \Theta) = \sum_{i=1}^N \sum_{g=1}^{m_i} \frac{u_{ig}}{P_{ig}} P'_{ig}(\Theta) \quad (3.4)$$

where $P'_{ig}(\Theta) = \frac{\partial P_{ig}(\Theta)}{\partial \theta_k}$. The explicit form for $P'_{ig}(\Theta)$ is given by

$$\begin{aligned} \frac{\partial P_{ig}(\Theta)}{\partial \theta_k} &= \frac{\partial P_{ig-1}^*(\Theta)}{\partial \theta_k} - \frac{\partial P_{ig}^*(\Theta)}{\partial \theta_k} \\ &= Da_{ki}[1 - P_{ig-1}(\Theta)]P_{ig-1}(\Theta) - Da_{ki}[1 - P_{ig}(\Theta)]P_{ig}(\Theta) \end{aligned} \quad (3.5)$$

Substituting (3.5) into (3.4) and simplifying, we have

$$\begin{aligned} \frac{\partial}{\partial \theta_k} \log L(U | \Theta) &= \sum_{i=1}^N \sum_{g=1}^{m_i} \frac{Da_{ki}[P_{ig-1}^*(1 - P_{ig-1}^*) - P_{ig}^*(1 - P_{ig}^*)]}{P_{ig-1}^* - P_{ig}^*} * u_{ig} \\ &= \sum_{i=1}^N \sum_{g=1}^{m_i} Da_{ki}[1 - P_{ig-1}^* - P_{ig}^*] * u_{ig} \end{aligned} \quad (3.6)$$

for $k = 1, 2, \dots, h$.

Since the likelihood Equation 2.11 has no closed form solution, an iterative numerical procedure must be used. One standard method is the Newton-Raphson procedure. Let $\Theta^{(j)}$ denote the j -th approximation to the value of Θ that maximizes $L(U|\Theta)$. Then, provided $\Theta^{(j)}$ is in the neighborhood of the maximum, an approximation with an even higher likelihood is given by

$$\Theta^{(j+1)} = \Theta^{(j)} - \delta^{(j)} \quad (3.7)$$

where $\delta^{(j)}$ is the $h \times 1$ vector

$$\delta^{(j)} = [H(\Theta^{(j)})]^{-1} \times \frac{\partial}{\partial \Theta} \log L(U | \Theta^{(j)}) \quad (3.8)$$

The matrix $H(\Theta^{(j)})$ is the $h \times h$ matrix of second derivatives evaluated at $\Theta^{(j)}$. The elements of $H(\Theta)$ can be expressed by the $h \times h$ symmetric matrix

$$H(\Theta) = \begin{bmatrix} \frac{\partial^2 \log L}{\partial^2 \theta_1} & \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_h} \\ & \frac{\partial^2 \log L}{\partial^2 \theta_2} & \dots & \frac{\partial^2 \log L}{\partial \theta_2 \partial \theta_h} \\ & & \ddots & \vdots \\ & & & \frac{\partial^2 \log L}{\partial^2 \theta_h} \end{bmatrix} \quad (3.9)$$

The diagonal elements of $H(\Theta)$ take the form

$$\begin{aligned} \frac{\partial^2}{\partial^2 \theta_k} \log L(U | \Theta) &= - \sum_{i=1}^N \sum_{g=1}^{m_i} u_{ig} \left(\frac{P'_{ig}}{P_{ig}} \right)^2 \\ &= - \sum_{i=1}^N \sum_{g=1}^{m_i} D^2 a_{ki}^2 (1 - P_{ig}^* - P_{ig+1}^*)^2 u_{ig} \end{aligned} \quad (3.10)$$

The off-diagonal elements are of the form

$$\frac{\partial^2}{\partial \theta_k \partial \theta_l} = - \sum_{i=1}^N \sum_{g=1}^{m_i} D^2 a_{ki} a_{li} (1 - P_{ig}^* - P_{ig+1}^*)^2 u_{ig} \quad (3.11)$$

In (3.8), $\frac{\partial}{\partial \Theta} \log L(U | \Theta^{(j)})$ is the $h \times 1$ vector partial derivatives (evaluated at $\Theta^{(j)}$)

defined by (3.2). Successive approximations are repeatedly obtained using (3.7) and (3.8) until the elements of $\delta^{(j)}$ become sufficiently small.

3.1.1.2 Bayesian-based method

Bayesian-based approaches to ability estimation differ from MLE in that assumptions about the nature of the population ability distribution are incorporated into the ability estimate. Instead of being based only on the likelihood function, estimates of abilities maximize the

natural logarithm of the posterior distribution. In educational testing, Θ usually follows a multivariate normal with mean vector X and covariate matrix Φ :

$$f(\Theta) = (2\pi)^{-h/2} |\Phi|^{-1/2} \exp\left[\frac{1}{2}(\Theta - X)' \Phi^{-1}(\Theta - X)\right] \quad (3.12)$$

According to Bayes theorem, the posterior density function of Θ is expressed by

$$f(\Theta|U) \propto L(U|\Theta)f(\Theta) \quad (3.13)$$

where $L(U|\Theta)$ is the likelihood function given by Equation 2.11, $f(\Theta)$ is the prior distribution of Θ , which is defined by Equation 3.12.

The modal estimates, denoted by $\hat{\Theta}$, are the values of Θ that satisfy the set of H simultaneous equations given by

$$\frac{\partial}{\partial \Theta} \log f(\Theta | U) = 0 \quad (3.14)$$

where

$$\frac{\partial}{\partial \Theta} \log f(\Theta | U) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \log f(\Theta | U) \\ \frac{\partial}{\partial \theta_2} \log f(\Theta | U) \\ \vdots \\ \frac{\partial}{\partial \theta_h} \log f(\Theta | U) \end{bmatrix} \quad (3.15)$$

Explicit expressions for these partial derivatives can be obtained by noting that the natural logarithm of the posterior density function is

$$\begin{aligned} \log f(\Theta|U) &= \log L(U|\Theta) + \log f(\Theta) \\ &= \log L(U|\Theta) - \frac{1}{2}(\Theta - X)' \Phi^{-1}(\Theta - X) \end{aligned} \quad (3.16)$$

Then we have

$$\frac{\partial \log f(\Theta | U)}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \log L(U | \Theta) - \frac{1}{2} \frac{\partial}{\partial \theta_k} [(\Theta - X)' \Phi^{-1}(\Theta - X)] \quad (3.17)$$

Expressions for the first term, $\frac{\partial \log L(U | \Theta)}{\partial \theta_k}$, are provided by Equation 3.6. The explicit expression for the second term takes the form

$$\frac{\partial}{\partial \theta_k} [(\Theta - X)' \Phi^{-1}(\Theta - X)] = 2 \times \left[\frac{\partial}{\partial \theta_k} (\Theta - X)' \right] \Phi^{-1}(\Theta - X) \quad (3.18)$$

where $\frac{\partial}{\partial \theta_k}(\Theta - X)'$ denotes a $1 \times h$ vector with the k -th element set equal to 1 and all other

elements equal to zero. Substituting Equation 3.6 and 3.18 into 3.17, we have

$$\frac{\partial \log f(\Theta | U)}{\partial \theta_k} = \sum_{i=1}^N \sum_{g=1}^{m_i} Da_{ki} [1 - P_{ig}^* - P_{ig+1}^*] - \left[\frac{\partial}{\partial \theta_k}(\Theta - X)' \right] \Phi^{-1}(\Theta - X) \quad (3.19)$$

for $k = 1, 2, \dots, h$.

As with the likelihood equation, the equation given by Equation 3.19 has no explicit solutions, so an iterative numerical procedure such as the Newton-Raphson procedure must be used. Accordingly, if we let $\Theta^{(j)}$ denote the j -th approximation to the value of Θ that maximizes $\log f(\Theta | U)$, then a better approximation is generally given by

$$\Theta^{(j+1)} = \Theta^{(j)} - \delta^{(j)} \quad (3.20)$$

where $\delta^{(j)}$ is the $h \times 1$ vector

$$\delta^{(j)} = [J(\Theta^{(j)})]^{-1} \times \frac{\partial}{\partial \Theta} \log f(\Theta^{(j)} | U) \quad (3.21)$$

The matrix $J(\Theta^{(j)})$ is the $h \times h$ matrix of second derivatives evaluated at $\Theta^{(j)}$. The elements of $J(\Theta)$ are expressed by the $h \times h$ symmetric matrix

$$J(\Theta) = \begin{bmatrix} \partial^2 \log f(\Theta | U) / \partial^2 \theta_1 & \partial^2 \log f(\Theta | U) / \partial \theta_1 \partial \theta_2 & \cdots & \partial^2 \log f(\Theta | U) / \partial \theta_1 \partial \theta_h \\ & \partial^2 \log f(\Theta | U) / \partial^2 \theta_2 & \cdots & \partial^2 \log f(\Theta | U) / \partial \theta_2 \partial \theta_h \\ & & \ddots & \vdots \\ & & & \partial^2 \log f(\Theta | U) / \partial^2 \theta_h \end{bmatrix}$$

Taking the derivative of Equation 3.19, we see that the diagonal elements of $J(\Theta)$ take the form

$$\frac{\partial^2}{\partial \theta_k^2} \log f(\Theta | U) = \frac{\partial^2}{\partial \theta_k^2} \log L(U | \Theta) - \frac{1}{2} \frac{\partial^2}{\partial \theta_k^2} [(\Theta - X)' \Phi^{-1}(\Theta - X)] \quad (3.22)$$

The first term on the right hand side of Equation 3.22 is given by Equation 3.10. The explicit expression for the second term is given by

$$\frac{\partial^2}{\partial \theta_k^2} [(\Theta - X)' \Phi^{-1}(\Theta - X)] = 2 \times \left[\frac{\partial}{\partial \theta_k}(\Theta - X)' \right] \Phi^{-1} \left[\frac{\partial}{\partial \theta_k}(\Theta - X) \right] \quad (3.23)$$

Substituting Equation 3.10 and 3.23 into Equation 3.22 we have

$$\frac{\partial^2 \log f(\Theta | U)}{\partial \theta_k^2} = -D^2 \sum_{i=1}^N \sum_{g=1}^{m_i} a_{ki}^2 (1 - P_{ig}^* - P_{ig+1}^*)^2 u_{ig} - \phi_{kk} \quad (3.24)$$

where ϕ_{kk} is the k -th diagonal element of Φ^{-1} . The off-diagonal elements of $J(\Theta)$ take the form

$$\frac{\partial^2}{\partial \theta_k \partial \theta_l} \log f(\Theta | U) = \frac{\partial^2}{\partial \theta_k \partial \theta_l} \log L(U | \Theta) - \frac{1}{2} \frac{\partial^2}{\partial \theta_k \partial \theta_l} [(\Theta - X)' \Phi^{-1} (\Theta - X)] \quad (3.25)$$

From Equation 3.11 and by evaluating the second term we have

$$\frac{\partial^2}{\partial \theta_k \partial \theta_l} \log f(\Theta | U) = -D^2 \sum_{i=1}^N \sum_{g=1}^{m_i} a_{ki} a_{li} (1 - P_{ig}^* - P_{ig+1}^*)^2 u_{ig} - \phi_{kl} \quad (3.26)$$

where ϕ_{kl} is the $\{k$ -th, l -th $\}$ element of Φ^{-1} . The vector of elements $\frac{\partial \log f(\Theta^{(j)} | U)}{\partial \Theta}$ in Equation

3.21 is the $h \times 1$ vector of partial derivatives (evaluated at $\Theta^{(j)}$) defined by Equation 3.19. Modal estimates can be obtained through successive approximations using Equation 3.20 and 3.21. Additional approximations are obtained until the elements of $\Theta^{(j)}$ change very little from one iteration to the next.

When the Newton-Raphson procedure is used, the first and second derivatives must be computed. Comparing Equation 3.6 with Equation 3.19, Equation 3.10 with Equation 3.24, and Equation 3.11 with Equation 3.26, it is apparent that the Bayesian-based method considers the correlation between different dimensions (Φ), while MLE doesn't.

3.1.2 Item selection methods

In adaptive testing, items are selected on the basis of item information. Suppose that $(i-1)$ items have been administered and $\hat{\Theta}$ is the ability estimate after $(i-1)$ responses, the task is to decide which item is to be administered as the i th item from the set of remaining items. For uni-dimensional CAT, the item information function is defined by

$$I(\hat{\theta}) = \sum_i \sum_{g=0}^{m_i} \frac{(P'_{ig})^2}{P_{ig}} \quad (3.27)$$

The item, which has the largest information value, is the one that should be selected. But for multi-dimensional case, the item information function is an $h \times h$ information matrix, denoted by $I(\hat{\Theta})$.

$$I(\hat{\Theta}) = \begin{bmatrix} \sum_i \sum_{g=1}^{m_i} \frac{(\partial P_{ig} / \partial \theta_1)^2}{P_{ig}} & \dots & \sum_i \sum_{g=1}^{m_i} \frac{\partial P_{ig} / \partial \theta_1 \times \partial P_{ig} / \partial \theta_h}{P_{ig}} \\ & \ddots & \vdots \\ & & \sum_i \sum_{g=1}^{m_i} \frac{(\partial P_{ig} / \partial \theta_h)^2}{P_{ig}} \end{bmatrix} \quad (3.28)$$

The diagonal elements of $I(\hat{\Theta})$ take the form

$$\begin{aligned} I_{rr}(\hat{\Theta}) &= \sum_i \sum_{g=1}^{m_i} \frac{(\partial P_{ig} / \partial \theta_r)^2}{P_{ig}} \\ &= \sum_i \sum_{g=1}^{m_i} D^2 a_{ri}^2 (1 - P_{ig}^* - P_{ig+1}^*)^2 (P_{ig}^* - P_{ig+1}^*) \end{aligned} \quad (3.29)$$

Similarly, the off-diagonal elements are

$$\begin{aligned} I_{rs}(\hat{\Theta}) &= \sum_i \sum_{g=1}^{m_i} \frac{(\partial P_{ig} / \partial \theta_r) \times (\partial P_{ig} / \partial \theta_s)}{P_{ig}} \\ &= \sum_i \sum_{g=1}^{m_i} D^2 a_{ri} a_{si} (1 - P_{ig}^* - P_{ig+1}^*)^2 (P_{ig}^* - P_{ig+1}^*) \end{aligned} \quad (3.30)$$

The problem now is to select an item based on this information matrix. Segall (1996) used the multivariate normal ellipsoid to provide a solution based on the dichotomous model. In this section, the item selection methods based on the multi-dimensional graded response model are discussed.

3.1.2.1 Maximum information

The distribution of the ability estimate $\hat{\Theta}_j$ (obtained from the first j responses) follows a multivariate normal distribution $N(\hat{\Theta}_0, \Sigma_j)$. A multivariate item-selection analog is motivated by the expression for the volume of the multivariate normal ellipsoid. Then

$$\text{Prob}[\hat{\Theta}_j' \sum_j^{-1} \hat{\Theta}_j \leq \chi_p^2(\alpha)] = 1 - \alpha \quad (3.31)$$

That is, the probability is $1 - \alpha$ that $\hat{\Theta}_j$ will fall inside the ellipsoid

$$x' \sum_j^{-1} x = \chi_p^2(\alpha) \quad (3.32)$$

The volume of this ellipsoid is

$$\varsigma \times |\sum_j|^{1/2} \quad (3.33)$$

where

$$\varsigma = \frac{2\pi^{p/2} [\chi_p^2(\alpha)]^{p/2}}{p\Gamma(\frac{1}{2}p)}, \quad (3.34)$$

and $\Gamma(\cdot)$ denotes the gamma function.

When considering items for administration, the multivariate analog to the univariate procedure selects the item that achieves the largest decrement in the volume of the confidence ellipsoid. We denote the volume decrement achieved by the administration of item k' by

$$V_{k'} = \varsigma |\sum_j|^{1/2} - \varsigma |\sum_{j+k'}|^{1/2} \quad (3.35)$$

where \sum_j is the dispersion matrix of the $p \times 1$ vector of provisional estimates $\hat{\Theta}_j$ obtained after the j -th response, and $\sum_{j+k'}$ is the dispersion matrix of provisional estimates obtained after administration of the first j items and the administration of item k' . \sum_j can be approximated by the inverse of the information matrix, given by

$$\sum_j = \{I(\Theta, \hat{\Theta}_j)\}^{-1} = [\sum_{i \in v} I(\Theta, U_i)]^{-1} \quad (3.36)$$

The covariance matrix of provisional estimates which includes the administration of item k' is given by

$$\sum_{j+k'} = [I(\Theta, \hat{\Theta}_j) + I(\Theta, U_{k'})]^{-1} \quad (3.37)$$

Substituting Equation 3.37 and 3.36 into Equation 3.35, we obtain

$$V_{k'} = \varsigma |I(\Theta, \hat{\Theta}_j)|^{-1/2} - \varsigma |[I(\Theta, \hat{\Theta}_j) + I(\Theta, U_{k'})]^{-1}|^{1/2} \quad (3.38)$$

Note that the first term is a constant across items, since ς depends only on p and α , and $|I(\Theta, \hat{\Theta}_j)|$ is based on previously administered items. Since in the second term ς remains constant over candidate items, $V_{k'}$ can be maximized by selecting the item that maximizes the quantity

$$|I(\Theta, \hat{\Theta}_j) + I(\Theta, U_{k'})| \quad (3.39)$$

3.1.2.2 Bayesian based method

As always, Bayesian methods use the posterior distribution instead of a likelihood function. For a normal posterior density function, the volume decrement achieved by the administration of item k' is given by

$$C_{k'} = \zeta |W_j^{-1}|^{1/2} - \zeta |W_{j+k'}^{-1}|^{1/2} \quad (3.40)$$

where W_j^{-1} is the covariance matrix of the posterior distribution computed from the first j items, $W_{j+k'}^{-1}$ is the covariance matrix incorporating $j + 1$ items (the first j items plus item k'), and ζ is defined by Equation 3.34. For the purpose of item selection, we approximate the posterior density function $f(\Theta|U)$ by a multivariate normal density with covariance matrix W^{-1} , where $W = -E[J(\Theta)]$,

$$(3.41)$$

and where $-E[J(\Theta)]$ is evaluated at the mode of the posterior distribution $\hat{\Theta}$. Taking the expectation of $J(\Theta)$, we see that the diagonal elements of W take the form

$$W_{rr} = \sum_i \sum_{g=1}^{m_i} D^2 a_{ri}^2 (1 - P_{ig}^* - P_{ig+1}^*)^2 (P_{ig}^* - P_{ig+1}^*) + \phi_{rr} \quad (3.42)$$

while the off-diagonal elements of W are

$$W_{rs} = \sum_i \sum_{g=1}^{m_i} D^2 a_{ri} a_{si} (1 - P_{ig}^* - P_{ig+1}^*)^2 (P_{ig}^* - P_{ig+1}^*) + \phi_{rs} \quad (3.43)$$

The matrix W_j is computed from Equation 3.42 and 3.43, where the summands are taken over the j adaptive administered items $v = \{v_1, v_2, \dots, v_j\}$, whereas the matrix $W_{j+k'}$ is computed from the summands $v = \{v_1, v_2, \dots, v_j, v_{k'}\}$.

The expression for the volume decrement can be simplified by noting that the determinant of the inverse of W is the reciprocal of the determinant (Searle, 1982):

$$C_{k'} = \zeta |W_j|^{-1/2} - \zeta |W_{j+k'}|^{-1/2} \quad (3.44)$$

Note that the first term is a constant across candidate items, since ζ depends only on p and α , and $|W_j|$ is based on previously administered items. The second term is a function of both ζ

and the determinate of the matrix $W_{j+k'}$. Since ς remains constant over candidate items, $C_{k'}$ can be maximized by selecting the item k' which maximizes $|W_{j+k'}|$. We can note the relation between $|W_{j+k'}|$ and the criterion used in the maximum information procedure (Equation 3.39) from the equation

$$|W_{j+k'}| = |I(\theta, \hat{\theta}_j) + I(\theta, u_{k'}) + \Phi^{-1}| \quad (3.45)$$

Note that the criterion for the maximum information item selection (Equation 3.39) and the criterion for the Bayesian item selection based on $|W_{j+k'}|$ differ only by the term which consists of the inverse of the covariance matrix of the prior distribution of abilities Φ^{-1} .

3.2 STUDY 1

3.2.1 Experimental design

The first study used simulated data and compared a three-dimension application of CAT to separate uni-dimensional CAT applications. The item pool consisted of items that were scored using five ordered categories. These item types are consistent with many psychological measures currently used. Two design factors were manipulated: 1) correlation between dimensions, and 2) item pool size.

As discussed, when the dimensions being measured are correlated, responses to items from one dimension provide information about an individual's status on the other dimensions (Segall, 2000). Thus, a multi-dimensional application of a CAT may exhibit efficiencies beyond current methods based on uni-dimensional models. However, the gain in any efficiency is likely to be dependent on the magnitude of the relationships between the traits. In order to explore this dependency, three levels of correlations were manipulated that reflect a range of possible values (0.0, 0.4, and 0.7). A correlation of 0.0 was included as a condition that is identical to a uni-dimensional application. Correlations of 0.4 and 0.7 were included to reflect medium and high correlations between dimensions.

In order to assess the effect of the size of item pool on the performance of the CAT, four item pools (10, 20, 50, 100 items for each dimension) were used. An item pool of 20 items reflects the length of many instruments that are currently used in psychological and health assessments. An item pool size of 10 reflects short forms of instruments that are typically used in psychological and health assessments to screen subjects. The 10 and 20 item pool sizes provide useful baseline conditions. In addition, in order to take advantage of CAT methodology, a large pool of items is necessary to address the full ability spectrum. Since more items near or around locations on the scale for a trait allow for more efficient measurement, exploration of the efficiencies associated with item pool size is relevant to the design of a CAT. Researchers have indicated that IRT methods allow for linking numerous instruments that measure different levels of a trait (McHorney & Cohen, 2000; Ware, Bjorner, and Kosinski, 2000). Thus, it is quite conceivable that item pools could reflect a large set of items. Item pools based on 50 and 100 items per instrument were also evaluated since these reflect large pools that are possible to achieve in practice (e.g., McHorney & Cohen, 2000).

For this study, item parameters for 10 items measuring each dimension were defined and replicated to achieve the different item pool sizes. Because a simple structure is used in this study, each item is allowed to possess one nonzero discrimination parameter. As an example, items in the first dimension are of the form $A_i' = \{a_{1i}, 0, 0\}$. Similarly, items in the second and third dimensions took the form $A_i' = \{0, a_{2i}, 0\}$ and $A_i' = \{0, 0, a_{3i}\}$ respectively. From Equation 3.29 and 3.30, item information is mainly dependent on the discrimination parameter a^2 (for simple structure, Equation 3.30 = 0). The item which has the largest discrimination parameter, usually provides the most item information. The discrimination parameters were fixed at three levels: 2.55, 1.70, and 0.85 to reflect high, medium, and low discrimination levels (given a scaling parameter (D) equal to 1.0).

For a high quality item pool, items should be evenly and equally distributed throughout the θ continuum of interest (Urry, 1977; Weiss, 1982). Based on estimated graded response model parameters from a real health assessment (DASH), a set of ten location parameters were carefully selected so that they were approximately equally distributed. The discrimination parameter 2.55 was assigned to items which are located at the average θ value. The discrimination parameter 1.70 and 0.85 were randomly assigned to items which were located at

the extreme θ value. This was consistent with the real data analysis. These ten item parameters form the first dimension of the item pool size = 10. Figure 3.1 shows the item information functions for these ten items. Table 3.1 presents the item parameters for the item pool size = 10. For other item pool sizes, a value between -0.5 and 0.5 was randomly sampled from a uniform distribution. This value was added to the above ten location parameters to shift the information function along the trait continuum. The range [-0.5, 0.5] was selected in order to avoid moving the items to extreme θ values. The three fixed discrimination parameters were assigned to items as before.

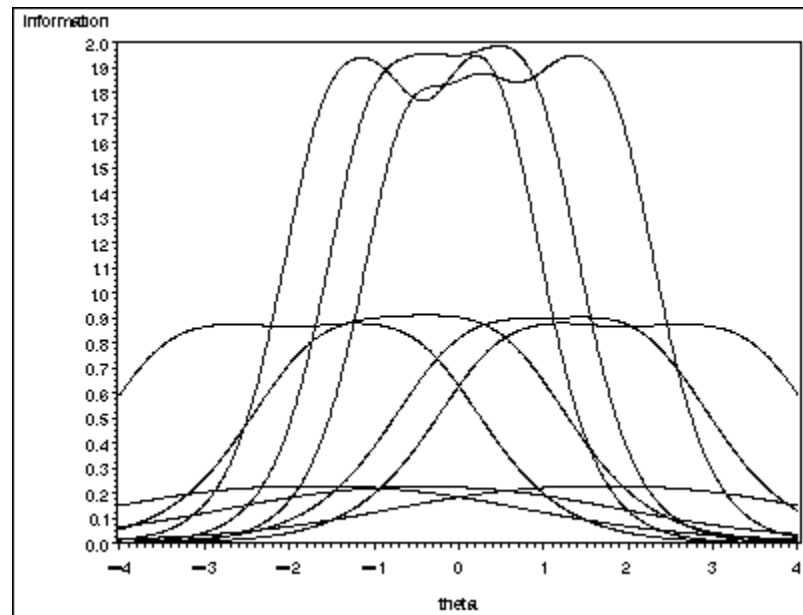


Figure 3.1. Item Information Curves for Pool Size = 10

A multi-dimensional CAT can administer several sub-tests. Examinees can receive unequal numbers of items from different content areas, which results in ability estimates based on different mixtures of content areas. As an example, the SF-36 health survey consists of items from two areas: physical and mental functions. Examinee A could receive 5 items from physical function and 10 items from mental function. But examinee B could receive 6 items from physical function and 8 items from mental function. In the present study, the number of items from each dimension was fixed for all examinees.

In all, there are 12 (3 correlations x 4 item pool sizes) combinations of conditions. The steps in simulating the CAT were as follows for each combination. First, 1000 vectors of true ability values from a multivariate normal distribution where the means for each dimension are 0 and the correlation defined by the simulation condition (0, 0.4, and 0.7) were simulated. A large number of vectors were simulated in order to have more stable estimates of the performance for low and high values of ability. The administration order follows the order of dimensions. As an example, for a 10-item test, 10 items were selected from the item pool measuring the first dimension, and then 10 items were selected from the item pool measuring the second dimension, and so on. For selecting the initial item, provisional ability estimates will be fixed at 0, the midpoint of the scale for each dimension. An item was selected that provides maximum information at this provisional estimate using the multi-dimensional item selection method described in Chapter 3. A response to the item was simulated given the true ability and item parameters, after which the provisional estimate for the ability was updated using Bayesian estimator (Chapter 3).

Table 3.1. Item Parameters for Pool Size = 10

Discrimination Parameter			Location Parameter				Dimension
a_1	a_2	a_3	b_1	b_2	b_3	b_4	
1.7	0	0	0.489	1.381	2.490	3.438	1
0.85	0	0	0.552	1.246	2.100	2.882	1
1.7	0	0	-0.014	0.774	1.648	2.224	1
2.55	0	0	-1.556	-0.912	0.082	0.478	1
2.55	0	0	-1.033	-0.346	0.383	0.922	1
2.55	0	0	-0.591	0.283	1.170	1.806	1
1.7	0	0	-1.704	-0.872	-0.114	0.540	1
0.85	0	0	-1.408	-0.742	-0.044	0.568	1
1.7	0	0	-3.438	-2.490	-1.381	-0.489	1
0.85	0	0	-2.882	-2.100	-1.246	-0.552	1
0	1.7	0	0.489	1.381	2.490	3.438	2
0	0.85	0	0.552	1.246	2.100	2.882	2
0	1.7	0	-0.014	0.774	1.648	2.224	2
0	2.55	0	-1.556	-0.912	0.082	0.478	2
0	2.55	0	-1.033	-0.346	0.383	0.922	2
0	2.55	0	-0.591	0.283	1.170	1.806	2
0	1.7	0	-1.704	-0.872	-0.114	0.540	2
0	0.85	0	-1.408	-0.742	-0.044	0.568	2
0	1.7	0	-3.438	-2.490	-1.381	-0.489	2
0	0.85	0	-2.882	-2.100	-1.246	-0.552	2
0	0	1.7	0.489	1.381	2.490	3.438	3
0	0	0.85	0.552	1.246	2.100	2.882	3
0	0	1.7	-0.014	0.774	1.648	2.224	3
0	0	2.55	-1.556	-0.912	0.082	0.478	3
0	0	2.55	-1.033	-0.346	0.383	0.922	3
0	0	2.55	-0.591	0.283	1.170	1.806	3
0	0	1.7	-1.704	-0.872	-0.114	0.540	3
0	0	0.85	-1.408	-0.742	-0.044	0.568	3
0	0	1.7	-3.438	-2.490	-1.381	-0.489	3
0	0	0.85	-2.882	-2.100	-1.246	-0.552	3

3.2.2 Outcome measures

The performance of the CAT was evaluated for fixed test lengths of 5, 10, 15, and 20 items. The outcome measures included the correlations between estimated ability and true ability, root mean squared error (RMSE), bias, and standard error for trait estimates.

Indices of bias and RMSE were computed at each combination based on the following formulas:

$$RMSE(\hat{\theta}_i) = \sqrt{\frac{1}{N} \sum_{j=1}^N (\hat{\theta}_{i,j} - \theta_i)^2} \quad (3.46)$$

$$Bias(\hat{\theta}_i) = \frac{1}{N} \sum_{j=1}^N (\hat{\theta}_{i,j} - \theta_i) \quad (3.47)$$

where $i = 1, 2, \dots, h$, and N is the number of replications. The number of replications in this study is the number of examinees. $\hat{\theta}$ is the estimated ability and θ is the true ability. Bias is unsigned and directional, where positive values denote over-estimation, negative values denote under-estimation, and zero indicates no bias. However, positive values and negative values could be canceled out. RMSE is the square root of the mean of the squared differences between the estimates and true values. It is always positive and reflects the absolute distance from true values. The closer the RMSE is to zero the better the accuracy.

The standard error (SE) of the $\hat{\theta}$ was computed at the test information function for each simulated examinee using Equation 3.48.

$$SE = \frac{1}{\sqrt{I}} \quad (3.48)$$

Two correlations between estimated and true abilities, Pearson and intra-class correlation, were also computed. The Pearson correlation is based only on rank order, whereas the intra-class correlation considers both rank order and equivalence of magnitude. The definition of intra-class correlation is given by:

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \quad (3.49)$$

where σ_s^2 is the variance between subject and σ_e^2 is the variance within subject.

In order to assess the effectiveness of the CAT at different levels of θ , results at defined levels of each ability ($\theta \leq -1$, $-1 < \theta < 0$, $0 \leq \theta < 1$, and $\theta \geq 1$) for the multi-dimensional CAT applications were examined separately and compared. The levels analyzed reflect low, moderate, and high levels of ability.

3.3 STUDY 2

A real data CAT simulation was also conducted. A dataset consisting of approximately 3,000 item responses to the DASH (Disabilities of the Arm, Shoulder, and Hand) and SF-36 (MOS 36-Item Short-Form Health Survey) from outpatients treated at Centers for Rehabilitation Services was used.

The DASH is a 30-item self-reported questionnaire used routinely in clinical practice to assess rehabilitation outcomes for individuals with a variety of upper extremity impairments. The DASH is designed to measure physical function and symptoms. It includes 21 physical function items, six symptom items, and three social/role function items. There are also two optional four item modules: one is intended for athletes/performing artists and the other is for general working populations. All items use a five-category Likert-type scale. The patients who took this questionnaire are in a heterogeneous population, which includes both males and females, people who place low, moderate, or high demands on their upper limbs during their daily lives, and people with a variety of upper-limb disorders.

The SF-36 is a general measure of health status that measures eight domains of health including physical functioning, role limitation due to physical problems, bodily pain, general health, vitality, social functioning, role limitations due to emotional problems, and mental health. These domains can be combined to produce physical and mental component scores. The SF-36 is a 36-item survey, containing 2 to 10 items for each domain. All items use Likert-type scales, the number of order categories for each item range from 2 to 6. The dataset consisted of responses from approximately equal numbers of males and females and an approximate mean age of 50 years.

Research has found that the DASH is approximately uni-dimensional and the application of the GR model is appropriate (Irrgang & Stone, 2004; Stone & Irrgang, 2004). Research for the

SF-36 has demonstrated two general dimensions: the physical and mental health dimensions (Haley, McHorney, & Ware, 1994; Ware, Kosinski, & Keller, 1994). Therefore, a multi-dimensional GR model is more appropriate. For the DASH, the item parameters and examinees' ability were calibrated using MULTILOG (Thissen, 1991), and were treated as the population item parameters and true ability.

For the SF-36, MULTILOG (Thissen, 1991) was applied separately to the physical and mental health dimensions to obtain item parameters and examinees' ability, which were treated as population item parameters and true abilities. The item parameters from the DASH and SF-36 create a three-dimension item pool. The empirically based correlations between the three dimensions were computed and used in this study.

Adaptive administrations with fixed test lengths of 5 and 10 items for each dimension were simulated as in Study 1. One difference between Study 1 and Study 2 was that in Study 2 actual item responses were used in the CAT simulations. Thus, when an item was selected during the course of the CAT for each examinee, an actual response for a particular item from an examinee's set of item responses was used as opposed to simulating a response to the item. This approach affords potentially more realism to the simulated adaptive administration.

As in Study 1, the number of items per instrument that were administered, correlations between ability estimates and true values, bias and RMSE for ability estimates, and SE were examined. Results at defined levels of each true ability ($\theta \leq -1$, $-1 < \theta < 0$, $0 \leq \theta < 1$, and $\theta \geq 1$) for the multi-dimensional and uni-dimensional CAT applications were examined separately and compared.

3.4 COMPUTER SIMULATION

3.4.1 Main procedures and subroutines

A SAS program (MCAT) was developed to simulate the CAT (See Appendix A for SAS code). This program works for both uni-dimensional and multi-dimensional models and both dichotomous and polytomous cases. MCAT follows the main procedures as below.

Step 1: Set up variable values, which will be used later, such as, number of dimensions, number of examinees, test length, correlations, number of categories, etc.

Step 2: Read in all the item parameters. The item parameters are either generated (Study 1) or calibrated from real data (Study 2). All item parameters for each item pool are saved in a parameter file. The format is as below.

<i>discrimination parameter</i>			<i>location parameter</i>				<i>dimension</i>	<i>#of categories</i>
1.30	0.0	0.0	1.679	2.523	3.723	4.396	1	5
0.0	0.8	0.0	-0.013	0.774	1.648	2.223	2	5
0.0	0.0	0.3	-1.092	-0.521	0.329	0.932	3	5

Step 3: Read the examinee item responses. There are two options. For study 1, simulated item responses were generated. This is described in Step 3a. Study 2 used real data and real item responses were read in. This is described in Step 3b.

Step 3a: Generate item responses. Using the IRT model and item parameters, the probability of each response for an item was calculated. A random number among (0, 1) is generated and compared to the probabilities for responding at each score response.

Step 3b: Read in item responses from real data. The real responses are saved in *.dat files.

Step 4: Simulate CAT. This is an iterative process, which includes Step 4a and Step 4b. This step was repeated “test length” times.

Step 4a: Select item. The item, which provides the largest item information was selected. When calculating item information, the estimated abilities were used. For the first item, the abilities were set to zero.

Step 4b: Estimate abilities. Based on the item responses, abilities on all dimensions are estimated at same time.

Step 5: Output. After all “test length” items are administered, the root mean squared error, bias, and test item information were calculated and recorded.

Figure 3.2 shows the work flow chart. The main subroutines implementing these procedures include “Response_Gen”, “CatLoop”, “ItemSelect”, and “AbilityEst”. See appendix A for program lists.

Response_Gen: This subroutine generates item responses. For the graded response model, Equation 2.6 is used to calculate response probabilities. (N-1) probabilities was

calculated for N-category items. For example, if there are 5 categories, P_1^* , P_2^* , P_3^* , and P_4^* were calculated, with $P_0^* = 1$ and $P_5^* = 0$. A random number p among $(0, 1)$ was then generated. This number (p) was compared with P_i^* and P_{i+1}^* , $i = 0, \dots, 4$. If $P_i^* > p > P_{i+1}^*$, the item response was i . Therefore, the generated item responses were $0, \dots, 4$.

CatLoop: This subroutine implements the iterative process. In this loop, the subroutines of “ItemSelect” and “AbilityEst” are called. The likelihood function (Equation 3.3), the first derivative functions (Equation 3.19), the second derivative functions (Equation 3.24 and 3.26) are also defined in this subroutine. In this loop, the RMSE, bias, and test information are also calculated and saved to files.

ItemSelect: Two item selection methods (MLE and Bayesian based method) are implemented in this subroutine. The information matrix (Equation 3.28) is calculated first. Then the item, which provides the largest determinant of information matrix, is selected.

AbilityEst: Two ability estimation methods (MLE and Bayesian based method) are implemented in this subroutine. The likelihood function, the first derivative function, and the second derivative function have been defined in subroutine “CatLoop”. A function “NLPNRA” is called to estimate ability levels.

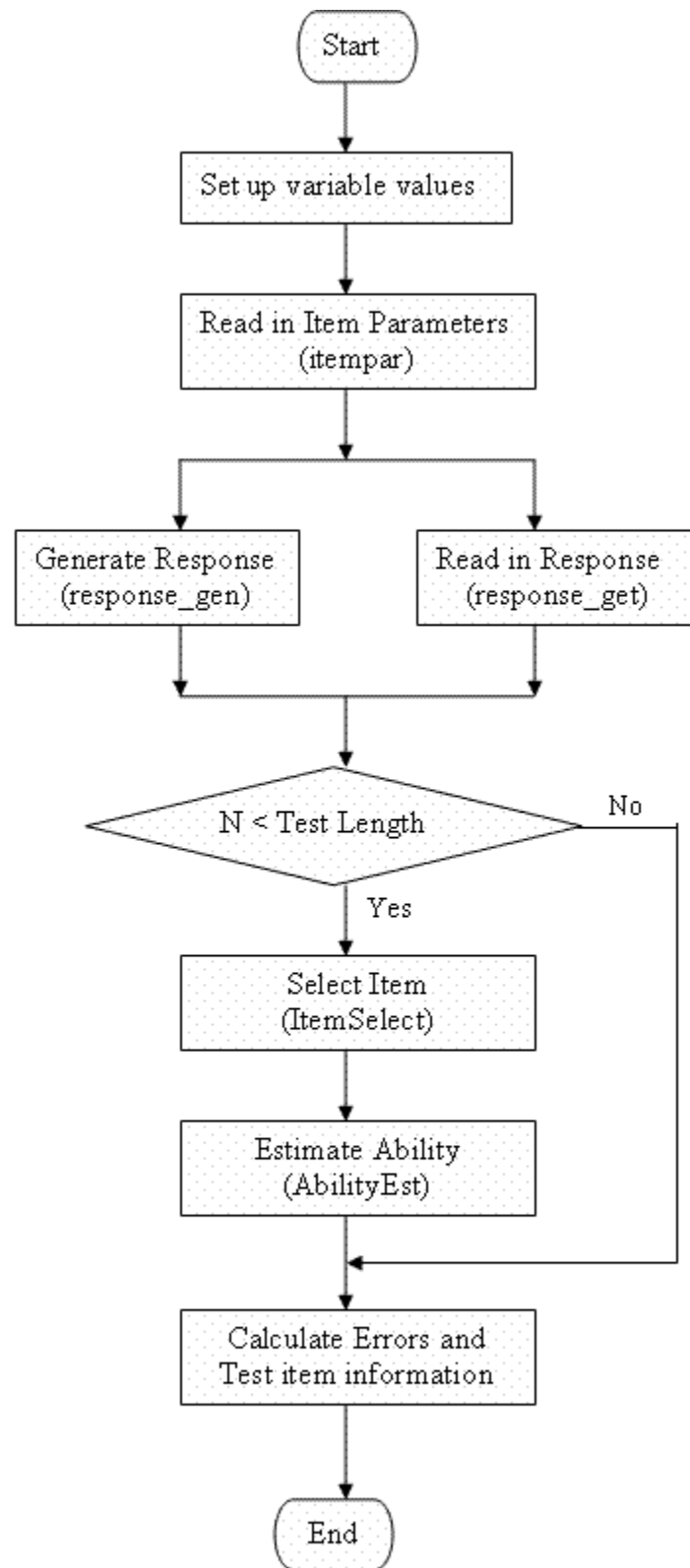


Figure 3.2. Flow Chart

3.4.2 Validation

In order to validate the MCAT program, the results from MCAT were compared with those from popular multi-dimensional computer programs, such as, NoHarm (Fraser, 1988) or TestFact (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, R., 2003), which may be used with only dichotomous items. Therefore, the multi-dimensional dichotomous case was used to check the generated item responses and ability estimation procedure of the program used in this study. When there are only two categories, the Samejima graded response model is the same as the dichotomous model. The results from the graded response model with two categories were compared with those based on the dichotomous model. In addition, the results from MCAT were compared with other researcher's work.

3.4.2.1 Validation of the generated item responses and ability estimation

1000 examinee item responses were generated by MCAT based on a one-dimension graded response model test. This test had 30 items, where each item had five categories. The item parameters were from the 10-item-pool and were repeated three times. These 30 item parameters are listed in Table 3.2 as the "true" column. True abilities were generated by MCAT and followed a standard normal distribution. MULTILOG was used to recover the item parameters based on graded response model. The item parameters which are estimated by MULTILOG are listed in Table 3.2 as "MULTILOG" column. Table 3.2 shows the item parameters are recovered very well.

Examinees' abilities were estimated by MCAT and MULTILOG separately. The true abilities and two estimated abilities from MCAT and MULTILOG were highly correlated, especially the correlation between estimated abilities from MCAT and MULTILOG. The scatter plots are shown in Figure 3.3 and the correlations between them were greater than or equal to .97:

	<i>True</i>	<i>MCAT</i>	<i>MULTILOG</i>
<i>True</i>	1.000	0.979	0.978
<i>MCAT</i>		1.000	0.999
<i>MULTILOG</i>			1.000

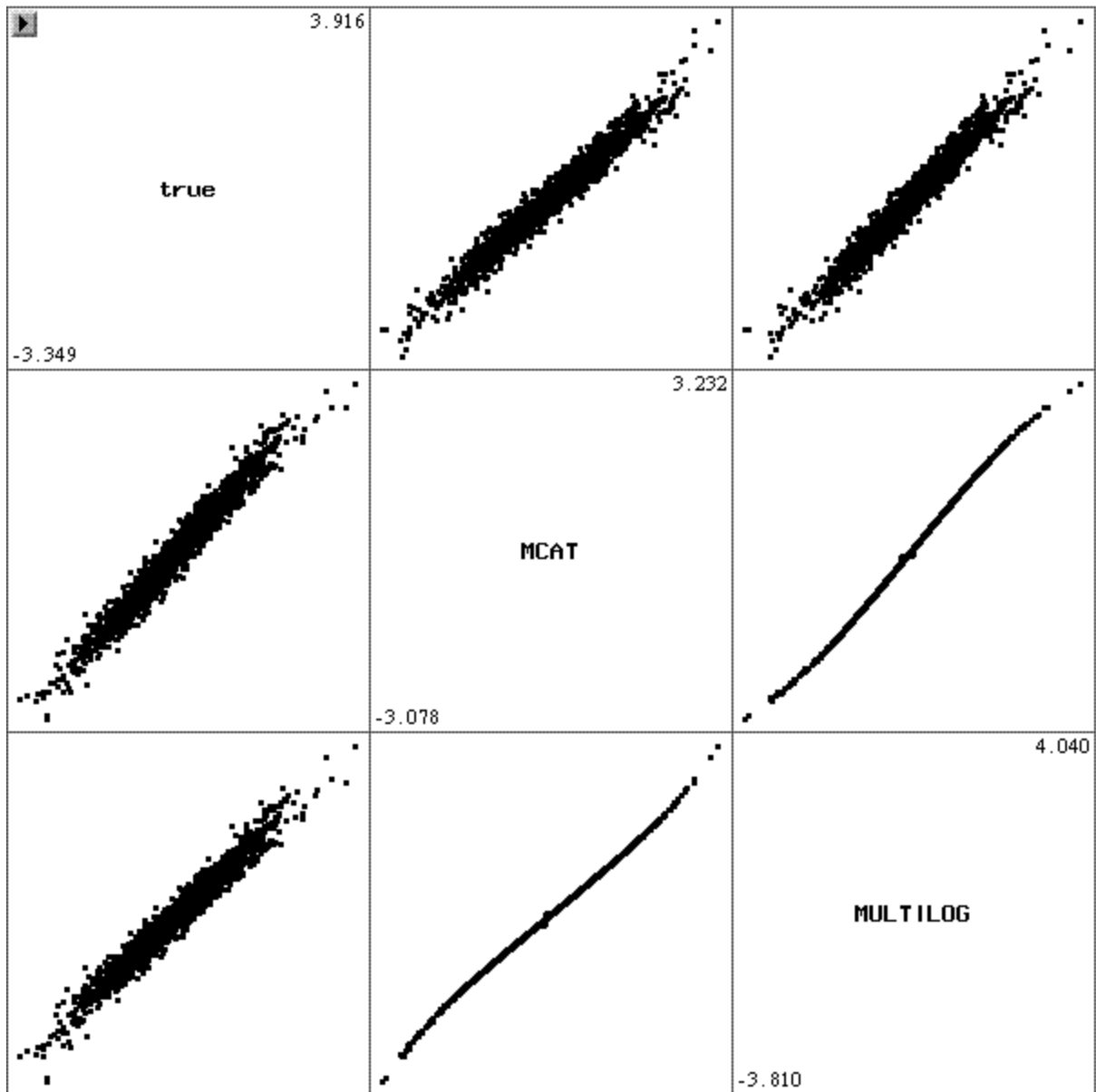


Figure 3.3. Scatter Plots

Table 3.2. One dimension: MCAT vs. MULTILOG

	A		B1		B2		B3		B4	
	TRUE	MULTILOG	TRUE	MULTILOG	TRUE	MULTILOG	TRUE	MULTILOG	TRUE	MULTILOG
1	1.70	1.54	0.489	0.569	1.381	1.51	2.49	2.8	3.438	4.13
2	0.85	0.833	0.552	0.531	1.246	1.25	2.1	2.15	2.882	2.9
3	1.70	1.58	-0.014	-0.0498	0.774	0.774	1.648	1.76	2.224	2.38
4	2.55	2.49	-1.556	-1.56	-0.912	-0.915	0.082	0.065	0.478	0.505
5	2.55	2.5	-1.033	-1.04	-0.346	-0.363	0.383	0.37	0.922	0.911
6	2.55	2.59	-0.591	-0.621	0.283	0.265	1.17	1.15	1.806	1.84
7	1.70	1.65	-1.704	-1.8	-0.872	-0.941	-0.114	-0.129	0.54	0.549
8	0.85	0.897	-1.408	-1.36	-0.742	-0.691	-0.044	-0.0965	0.568	0.536
9	1.70	1.69	-3.438	-3.42	-2.49	-2.5	-1.381	-1.4	-0.489	-0.508
10	0.85	0.87	-2.882	-2.69	-2.1	-2	-1.246	-1.22	-0.552	-0.547
11	1.70	1.68	0.489	0.459	1.381	1.38	2.49	2.56	3.438	3.59
12	0.85	0.781	0.552	0.606	1.246	1.36	2.1	2.25	2.882	3.13
13	1.70	1.69	-0.014	-0.0359	0.774	0.79	1.648	1.69	2.224	2.28
14	2.55	2.56	-1.556	-1.66	-0.912	-0.955	0.082	0.0863	0.478	0.455
15	2.55	2.52	-1.033	-1.08	-0.346	-0.37	0.383	0.408	0.922	0.934
16	2.55	2.45	-0.591	-0.673	0.283	0.286	1.17	1.16	1.806	1.8
17	1.70	1.61	-1.704	-1.81	-0.872	-0.969	-0.114	-0.145	0.54	0.561
18	0.85	0.881	-1.408	-1.42	-0.742	-0.761	-0.044	-0.0525	0.568	0.533
19	1.70	1.75	-3.438	-3.4	-2.49	-2.49	-1.381	-1.39	-0.489	-0.461
20	0.85	0.839	-2.882	-3.05	-2.1	-2.18	-1.246	-1.31	-0.552	-0.575
21	1.70	1.77	0.489	0.479	1.381	1.4	2.49	2.44	3.438	3.33
22	0.85	0.779	0.552	0.599	1.246	1.42	2.1	2.46	2.882	3.15
23	1.70	1.5	-0.014	-0.0108	0.774	0.856	1.648	1.88	2.224	2.56
24	2.55	2.46	-1.556	-1.54	-0.912	-0.923	0.082	0.082	0.478	0.511
25	2.55	2.38	-1.033	-1.06	-0.346	-0.369	0.383	0.352	0.922	0.977
26	2.55	2.38	-0.591	-0.602	0.283	0.269	1.17	1.23	1.806	1.9
27	1.70	1.64	-1.704	-1.76	-0.872	-0.86	-0.114	-0.134	0.54	0.527
28	0.85	0.833	-1.408	-1.4	-0.742	-0.713	-0.044	-0.0915	0.568	0.51
29	1.70	1.65	-3.438	-3.36	-2.49	-2.49	-1.381	-1.45	-0.489	-0.51
30	0.85	0.817	-2.882	-3.02	-2.1	-2.19	-1.246	-1.3	-0.552	-0.701

In addition, another 1000 examinee item responses were generated by MCAT based on a three-dimension test. This test had 20 items for each dimension, where each item had two categories. These 60 item parameters are listed in Table 3.3 as the “true” column. True abilities follow a multivariate normal distribution, where the means for each dimension were 0 and the correlations between all three dimensions were 0.6. NoHarm was used to recover the item parameters for this three-dimensional test. The estimated item parameters by NoHarm are listed

in Table 3.3 as column “NoHarm”. The correlations between three dimensions which were estimated by NoHarm were very close to 0.6 (.593, .597, .604).

Table 3.3. Three dimensions: MCAT vs. NoHarm

	A1		A2		A3		B	
	TRUE	NoHarm	TRUE	NoHarm	TRUE	NoHarm	TRUE	NoHarm
1	1.7	1.506	0	0	0	0	1.381	1.525
2	0.85	0.704	0	0	0	0	1.246	1.512
3	1.7	1.754	0	0	0	0	0.774	0.712
4	2.55	2.795	0	0	0	0	-0.912	-0.838
5	2.55	2.861	0	0	0	0	-0.346	-0.348
6	2.55	2.489	0	0	0	0	0.283	0.252
7	1.7	1.528	0	0	0	0	-0.872	-0.850
8	0.85	0.877	0	0	0	0	-0.742	-0.669
9	1.7	1.394	0	0	0	0	-2.49	-2.780
10	0.85	0.852	0	0	0	0	-2.1	-1.994
11	0	0	1.7	1.399	0	0	1.381	1.561
12	0	0	0.85	0.869	0	0	1.246	1.159
13	0	0	1.7	1.532	0	0	0.774	0.754
14	0	0	2.55	2.599	0	0	-0.912	-0.919
15	0	0	2.55	2.681	0	0	-0.346	-0.325
16	0	0	2.55	2.465	0	0	0.283	0.278
17	0	0	1.7	1.681	0	0	-0.872	-0.814
18	0	0	0.85	0.964	0	0	-0.742	-0.646
19	0	0	1.7	1.375	0	0	-2.49	-2.784
20	0	0	0.85	0.865	0	0	-2.1	-2.169
21	0	0	0	0	1.7	1.692	1.381	1.278
22	0	0	0	0	0.85	0.911	1.246	1.144
23	0	0	0	0	1.7	1.741	0.774	0.786
24	0	0	0	0	2.55	2.472	-0.912	-0.899
25	0	0	0	0	2.55	2.489	-0.346	-0.349
26	0	0	0	0	2.55	2.281	0.283	0.373
27	0	0	0	0	1.7	1.688	-0.872	-0.815
28	0	0	0	0	0.85	0.864	-0.742	-0.750
29	0	0	0	0	1.7	0.974	-2.49	-3.658
30	0	0	0	0	0.85	0.699	-2.1	-2.428

Examinees’ abilities were estimated by MCAT and TestFact separately. The true abilities and two estimated abilities from MCAT and TestFact were highly correlated, especially the correlations between estimated abilities from MCAT and TestFact. The correlations between them are shown as following:

	<i>True1</i>	<i>True2</i>	<i>True3</i>	<i>Est1</i>	<i>Est2</i>	<i>Est3</i>	<i>TestFact1</i>	<i>TestFact2</i>	<i>TestFact3</i>
<i>True1</i>	1.00			0.89			0.86		
<i>True2</i>		1.00			0.90			0.86	
<i>True3</i>			1.00			0.89			0.86
<i>Est1</i>				1.00			0.97		
<i>Est2</i>					1.00			0.97	
<i>Est3</i>						1.00			0.97
<i>TestFact1</i>							1.00		
<i>TestFact2</i>								1.00	
<i>TestFact3</i>									1.00

Note: True1, True2, and True3 are true abilities value. Est1, Est2, and Est2 are results from MCAT. TestFact1, TestFact2, and TestFact3 are results from TestFact.

3.4.2.2 Comparison with previous research

Segall (1996) conducted a simulation study to compare reliability values for multi-dimensional Bayesian ability estimates with their uni-dimensional counterparts. The simulated tests were based on the nine adaptive power tests of the CAT-ASVAB. The correlations between the nine dimensions are given in Table 3.4. A total of 15 conditions were simulated by Segall (1996). Test lengths for three of them are provided in Table 3.5. For each condition, the squared correlation between the true abilities and final modal estimates (reliability estimates) were calculated (see Table 3.6).

A nine-dimension test based on 2P model was simulated by MCAT using the same item parameters for the simulated study. 1000 examinee item responses were generated by MCAT based on 2P model. The correlations between nine dimensions were fixed at 0.6. Three test lengths (9, 18, and 27) were simulated. For each dimension, the test lengths were 1, 2, and 3 separately. The pool size was fixed at 100 items for each dimension. The reliability (the squared correlation) between the true abilities and estimated abilities were calculated and provided in Table 3.7. Although different item pools and correlations between true abilities were used in these two studies, the results in Tables 3.6 and 3.7 show similar results. Table 3.7 shows that MCAT produced a similar range in reliability values as that reported by Segall.

Table 3.4. Correlations between nine dimensions

	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
GS	1.000								
AR	.645	1.000							
WK	.908	.611	1.000						
PC	.808	.847	.880	1.000					
AI	.486	.332	.326	.349	1.000				
SI	.676	.424	.566	.514	.824	1.000			
MK	.564	.846	.516	.711	.150	.218	1.000		
MC	.739	.758	.644	.800	.623	.725	.625	1.000	
EI	.808	.639	.724	.743	.642	.724	.536	.822	1.000

Table 3.5. Test Lengths

	Number of Administered Items								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
MAT-9	1	1	1	1	1	1	1	1	1
MAT-17	2	2	2	1	2	1	3	2	2
MAT-24	2	3	3	2	2	2	4	3	3

Table 3.6. Reliability

	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
MAT-9	.719	.675	.676	.712	.588	.645	.624	.685	.679
MAT-17	.800	.785	.762	.798	.683	.717	.791	.780	.776
MAT-24	.837	.844	.818	.865	.702	.770	.840	.826	.826

Table 3.7. Reliability from MCAT

	Dimensions								
	1	2	3	4	5	6	7	8	9
9	0.632	0.598	0.613	0.632	0.623	0.646	0.642	0.63	0.626
18	0.726	0.729	0.736	0.729	0.74	0.745	0.736	0.752	0.748
27	0.781	0.792	0.797	0.806	0.808	0.801	0.799	0.884	0.52

4.0 RESULTS

The findings of the study will be presented in this section. In order to compare the efficiency of multi-dimensional CAT versus uni-dimensional CAT based on the graded response model, two studies were designed. One was based on simulated data, the other based on real data. Five factors that reflect realistic testing situations and that could affect the efficiency of computer adaptive testing were considered: (1) correlations between dimensions, (2) item pool size, (3) test length, (4) ability levels, and (5) number of dimensions used for trait estimation. The comparison was based on five outcome measures, including the Pearson and intra-class correlations between estimated ability and true ability, root mean squared error (RMSE), bias, and standard error of estimates. The real data CAT simulation was based on the DASH and SF-36 surveys. The correlations between the dimensions were analyzed. Five outcome measures were also calculated at defined levels of ability.

The results from these two studies are summarized and presented separately. For the simulation study, results are presented in six sections and describe the impact of the correlation, test length, ability level, number of dimensions used for trait estimation, and item pool size on the five outcome measures. The optimal size of the item pool for multi-dimensional CAT is also analyzed. For the real data study, the data analysis steps are described, as well as the impact of the correlation between dimensions, test length, and ability level on the outcome measures.

4.1 RESULTS FROM SIMULATED DATA

The simulated data are from a three-dimension application of MCAT. Two factors were directly manipulated: the correlations between dimensions and item pool size. In addition, five outcome measures were calculated for four fixed test lengths and for four ability levels: RMSE,

bias, Pearson correlation and intra-class correlation between estimated and true abilities, and standard error of estimates. Since the MCAT was simulated on each dimension separately and in succession, it was possible that the ability estimation in one dimension affected the ability estimation in subsequent dimensions. Therefore, the influence of five factors on the efficiency of MCAT was evaluated: item pool size, correlations between dimensions, test length, ability level, and the number of dimensions used for trait estimation. The following sections will present the results for each outcome measure.

4.1.1 Root Mean Squared Error (RMSE) Measure

The root mean squared error (RMSE – see Equation 3.46) for each combination (3 correlations between dimensions x 4 test length x 4 item pool size x 4 ability level) and for each dimension are presented in Table 4.1. Figures 4.1 through 4.4 illustrate the effect on the RMSE for each combination under different correlations, test lengths, item pool sizes, ability levels, and number of dimensions used for trait estimation respectively. For each figure there are three plots, one for each dimension. The “n” on the horizontal axes is a label for each combination of conditions. These may be translated using Table 4.2. It should be noted that there are blank cells in the table. These cells correspond to combinations of conditions that are missing by design. Test lengths of 15 and 20 items were not possible for an item pool size of 10 items.

Table 4.1. Root Mean Squared Error

Pool Size	Ability Level	Correlation	Dimension											
			1				2				3			
			Test Length				Test Length				Test Length			
			5	10	15	20	5	10	15	20	5	10	15	20
10	1	0	0.25	0.19			0.22	0.19			0.21	0.18		
		0.4	0.21	0.16			0.18	0.16			0.17	0.15		
		0.7	0.20	0.16			0.17	0.15			0.17	0.15		
	2	0	0.11	0.10			0.13	0.11			0.13	0.11		
		0.4	0.10	0.09			0.12	0.11			0.12	0.10		
		0.7	0.10	0.09			0.11	0.10			0.11	0.10		
	3	0	0.10	0.09			0.12	0.10			0.11	0.10		
		0.4	0.10	0.09			0.12	0.09			0.11	0.09		
		0.7	0.09	0.08			0.11	0.09			0.10	0.09		
	4	0	0.20	0.16			0.18	0.16			0.17	0.15		
		0.4	0.16	0.13			0.15	0.13			0.15	0.13		
		0.7	0.15	0.12			0.15	0.14			0.16	0.14		
	Total	0	0.38	0.35			0.39	0.36			0.38	0.35		
		0.4	0.36	0.33			0.37	0.34			0.36	0.34		
		0.7	0.35	0.32			0.36	0.33			0.35	0.33		
20	1	0	0.20	0.13	0.11	0.10	0.16	0.11	0.11	0.10	0.15	0.10	0.08	0.08
		0.4	0.17	0.11	0.09	0.09	0.15	0.1	0.09	0.09	0.12	0.09	0.08	0.07
		0.7	0.16	0.11	0.09	0.09	0.15	0.10	0.09	0.09	0.12	0.08	0.07	0.07
	2	0	0.09	0.06	0.05	0.05	0.10	0.06	0.05	0.05	0.09	0.06	0.06	0.06
		0.4	0.09	0.06	0.05	0.05	0.10	0.06	0.05	0.05	0.08	0.06	0.06	0.05
		0.7	0.08	0.05	0.05	0.05	0.09	0.06	0.05	0.05	0.08	0.06	0.05	0.05
	3	0	0.11	0.07	0.06	0.05	0.08	0.06	0.05	0.05	0.10	0.07	0.06	0.06
		0.4	0.1	0.07	0.05	0.05	0.08	0.06	0.05	0.05	0.10	0.07	0.06	0.06
		0.7	0.10	0.07	0.05	0.05	0.08	0.06	0.05	0.05	0.09	0.07	0.05	0.05
	4	0	0.18	0.11	0.09	0.09	0.13	0.09	0.08	0.07	0.16	0.10	0.09	0.09
		0.4	0.15	0.10	0.09	0.08	0.11	0.08	0.07	0.07	0.13	0.09	0.09	0.08
		0.7	0.15	0.10	0.08	0.08	0.10	0.08	0.07	0.07	0.13	0.09	0.09	0.08
	Total	0	0.35	0.28	0.26	0.25	0.33	0.27	0.26	0.25	0.34	0.28	0.26	0.26
		0.4	0.34	0.27	0.25	0.25	0.32	0.27	0.25	0.24	0.32	0.27	0.25	0.25
		0.7	0.33	0.27	0.25	0.24	0.31	0.26	0.25	0.24	0.31	0.27	0.25	0.25

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,
ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.
Total means: all ability levels.

Table 4.1. Continued

Pool Size	Ability Level	Correlation	Dimension											
			1				2				3			
			Test Length				Test Length				Test Length			
			5	10	15	20	5	10	15	20	5	10	15	20
50	1	0	0.17	0.10	0.08	0.07	0.12	0.07	0.05	0.04	0.12	0.06	0.05	0.04
		0.4	0.15	0.09	0.07	0.06	0.11	0.07	0.05	0.04	0.11	0.06	0.04	0.04
		0.7	0.15	0.10	0.08	0.06	0.11	0.07	0.05	0.04	0.11	0.06	0.04	0.04
	2	0	0.09	0.05	0.04	0.04	0.10	0.05	0.03	0.03	0.09	0.05	0.04	0.03
		0.4	0.08	0.05	0.04	0.03	0.09	0.05	0.03	0.03	0.09	0.05	0.04	0.03
		0.7	0.08	0.05	0.04	0.03	0.08	0.05	0.03	0.03	0.09	0.05	0.04	0.03
	3	0	0.09	0.05	0.03	0.03	0.09	0.05	0.04	0.03	0.10	0.06	0.04	0.04
		0.4	0.08	0.04	0.03	0.03	0.09	0.05	0.04	0.03	0.10	0.05	0.04	0.03
		0.7	0.08	0.04	0.03	0.03	0.08	0.05	0.03	0.03	0.09	0.05	0.04	0.03
	4	0	0.14	0.09	0.06	0.06	0.16	0.07	0.05	0.04	0.14	0.07	0.05	0.05
		0.4	0.13	0.09	0.06	0.06	0.12	0.06	0.05	0.04	0.12	0.06	0.05	0.04
		0.7	0.14	0.09	0.07	0.06	0.12	0.05	0.05	0.04	0.12	0.06	0.05	0.04
	Total	0	0.33	0.25	0.22	0.21	0.33	0.24	0.20	0.18	0.33	0.24	0.21	0.19
		0.4	0.32	0.24	0.21	0.20	0.31	0.23	0.20	0.18	0.32	0.23	0.20	0.19
		0.7	0.32	0.25	0.21	0.20	0.30	0.23	0.20	0.18	0.31	0.23	0.20	0.19
100	1	0	0.18	0.08	0.07	0.05	0.13	0.07	0.06	0.05	0.10	0.06	0.04	0.03
		0.4	0.15	0.07	0.06	0.05	0.12	0.07	0.06	0.05	0.10	0.05	0.04	0.03
		0.7	0.14	0.08	0.06	0.05	0.11	0.06	0.05	0.05	0.11	0.05	0.04	0.03
	2	0	0.09	0.05	0.03	0.03	0.10	0.05	0.03	0.02	0.09	0.04	0.03	0.02
		0.4	0.08	0.04	0.03	0.03	0.09	0.05	0.03	0.02	0.08	0.05	0.03	0.02
		0.7	0.07	0.04	0.03	0.03	0.09	0.05	0.03	0.02	0.08	0.05	0.03	0.02
	3	0	0.10	0.05	0.03	0.03	0.11	0.06	0.03	0.03	0.09	0.05	0.04	0.03
		0.4	0.09	0.05	0.03	0.03	0.09	0.05	0.03	0.03	0.09	0.05	0.03	0.03
		0.7	0.09	0.05	0.03	0.03	0.08	0.05	0.03	0.03	0.09	0.05	0.03	0.03
	4	0	0.16	0.08	0.06	0.05	0.13	0.07	0.05	0.04	0.13	0.06	0.04	0.03
		0.4	0.14	0.07	0.05	0.04	0.11	0.06	0.04	0.03	0.09	0.05	0.03	0.03
		0.7	0.14	0.08	0.06	0.05	0.10	0.06	0.04	0.03	0.10	0.06	0.03	0.03
	Total	0	0.34	0.24	0.21	0.18	0.34	0.24	0.20	0.18	0.32	0.22	0.19	0.16
		0.4	0.32	0.23	0.20	0.18	0.32	0.23	0.19	0.17	0.29	0.22	0.18	0.16
		0.7	0.32	0.24	0.2	0.18	0.30	0.23	0.19	0.17	0.30	0.22	0.19	0.16

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,
ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.
Total means: all ability levels.

Table 4.2. Interpretation of n in Figure 1 ~ Figure 4

N	Figure1			Figure2			Figure3			Figure4		
	pool	ability	length	pool	corr	length	pool	ability	corr	ability	corr	length
1	10	1	5	10	0	5	10	1	0	1	0	5
2	10	1	10	10	0	10	10	1	0.4	1	0	10
3	10	1	15	10	0	15	10	1	0.7	1	0	15
4	10	1	20	10	0	20	10	2	0	1	0	20
5	10	2	5	10	0.4	5	10	2	0.4	1	0.4	5
6	10	2	10	10	0.4	10	10	2	0.7	1	0.4	10
7	10	2	15	10	0.4	15	10	3	0	1	0.4	15
8	10	2	20	10	0.4	20	10	3	0.4	1	0.4	20
9	10	3	5	10	0.7	5	10	3	0.7	1	0.7	5
10	10	3	10	10	0.7	10	10	4	0	1	0.7	10
11	10	3	15	10	0.7	15	10	4	0.4	1	0.7	15
12	10	3	20	10	0.7	20	10	4	0.7	1	0.7	20
13	10	4	5	20	0	5	20	1	0	2	0	5
14	10	4	10	20	0	10	20	1	0.4	2	0	10
15	10	4	15	20	0	15	20	1	0.7	2	0	15
16	10	4	20	20	0	20	20	2	0	2	0	20
17	20	1	5	20	0.4	5	20	2	0.4	2	0.4	5
18	20	1	10	20	0.4	10	20	2	0.7	2	0.4	10
19	20	1	15	20	0.4	15	20	3	0	2	0.4	15
20	20	1	20	20	0.4	20	20	3	0.4	2	0.4	20
21	20	2	5	20	0.7	5	20	3	0.7	2	0.7	5
22	20	2	10	20	0.7	10	20	4	0	2	0.7	10
23	20	2	15	20	0.7	15	20	4	0.4	2	0.7	15
24	20	2	20	20	0.7	20	20	4	0.7	2	0.7	20
25	20	3	5	50	0	5	50	1	0	3	0	5
26	20	3	10	50	0	10	50	1	0.4	3	0	10
27	20	3	15	50	0	15	50	1	0.7	3	0	15
28	20	3	20	50	0	20	50	2	0	3	0	20
29	20	4	5	50	0.4	5	50	2	0.4	3	0.4	5
30	20	4	10	50	0.4	10	50	2	0.7	3	0.4	10
31	20	4	15	50	0.4	15	50	3	0	3	0.4	15
32	20	4	20	50	0.4	20	50	3	0.4	3	0.4	20
33	50	1	5	50	0.7	5	50	3	0.7	3	0.7	5
34	50	1	10	50	0.7	10	50	4	0	3	0.7	10
35	50	1	15	50	0.7	15	50	4	0.4	3	0.7	15
36	50	1	20	50	0.7	20	50	4	0.7	3	0.7	20
37	50	2	5	100	0	5	100	1	0	4	0	5
38	50	2	10	100	0	10	100	1	0.4	4	0	10
39	50	2	15	100	0	15	100	1	0.7	4	0	15
40	50	2	20	100	0	20	100	2	0	4	0	20
41	50	3	5	100	0.4	5	100	2	0.4	4	0.4	5
42	50	3	10	100	0.4	10	100	2	0.7	4	0.4	10
43	50	3	15	100	0.4	15	100	3	0	4	0.4	15
44	50	3	20	100	0.4	20	100	3	0.4	4	0.4	20
45	50	4	5	100	0.7	5	100	3	0.7	4	0.7	5
46	50	4	10	100	0.7	10	100	4	0	4	0.7	10
47	50	4	15	100	0.7	15	100	4	0.4	4	0.7	15
48	50	4	20	100	0.7	20	100	4	0.7	4	0.7	20
49	100	1	5									
50	100	1	10									
51	100	1	15									
52	100	1	20									
53	100	2	5									
54	100	2	10									
55	100	2	15									
56	100	2	20									
57	100	3	5									
58	100	3	10									
59	100	3	15									
60	100	3	20									
61	100	4	5									
62	100	4	10									
63	100	4	15									
64	100	4	20									

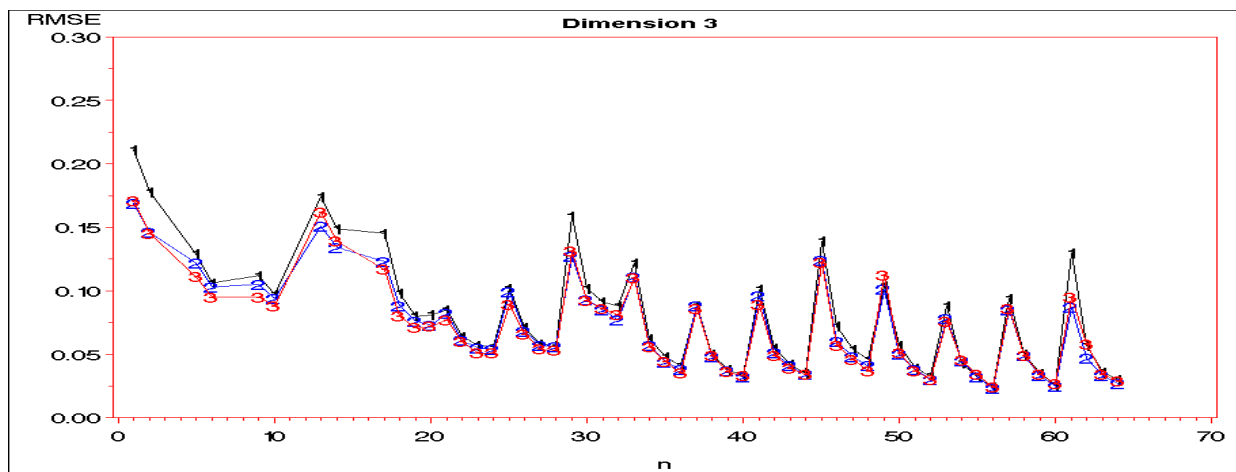
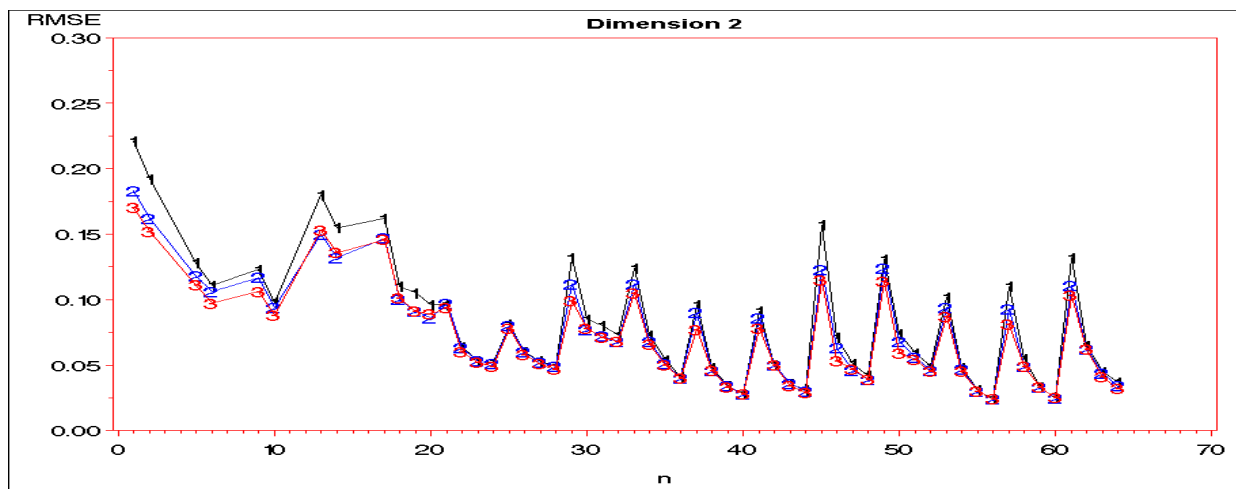
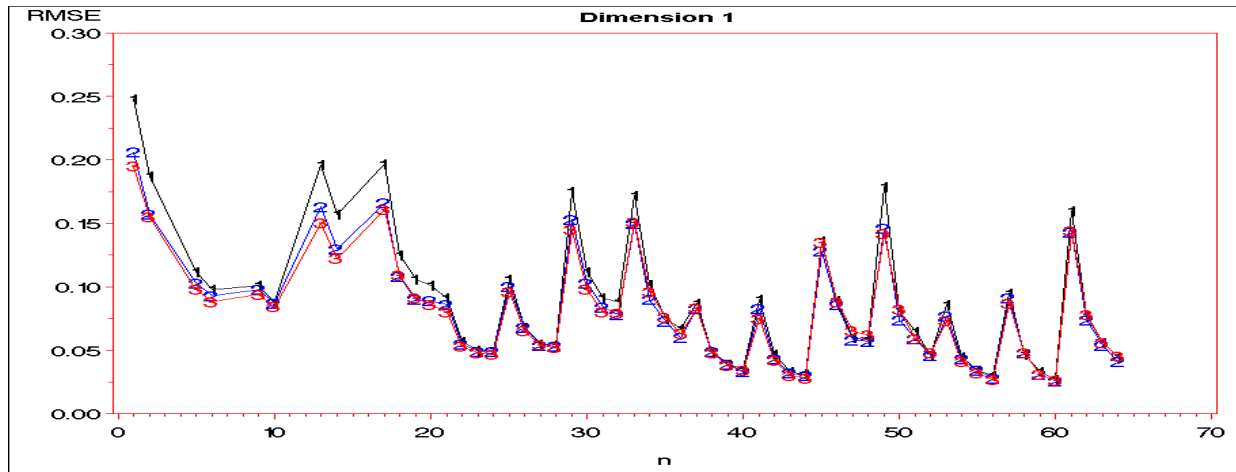


Figure 4.1. Root Mean Squared Error under different correlation
(line1: correlation = 0.0, line2: correlation = 0.4, line3: correlation = 0.7)

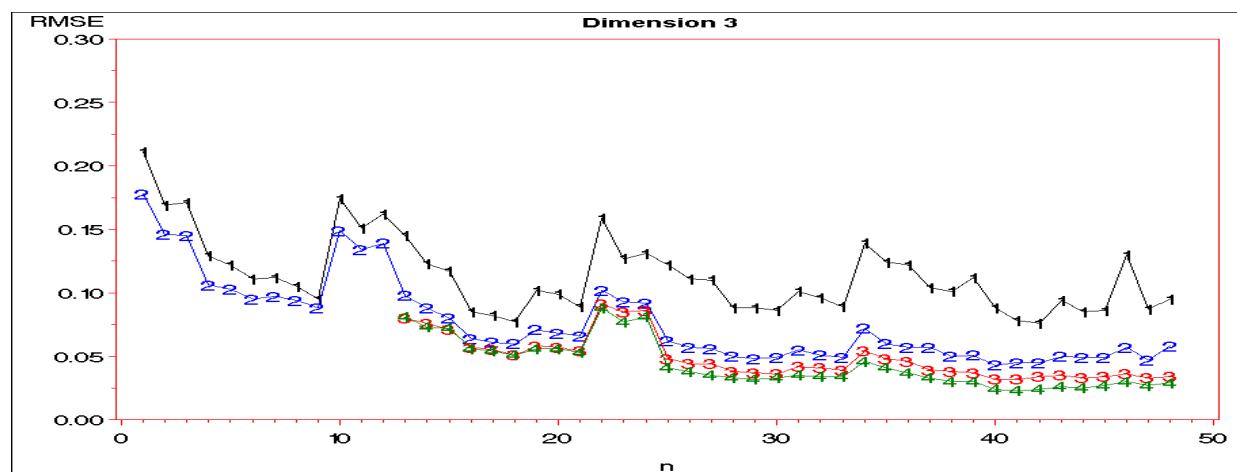
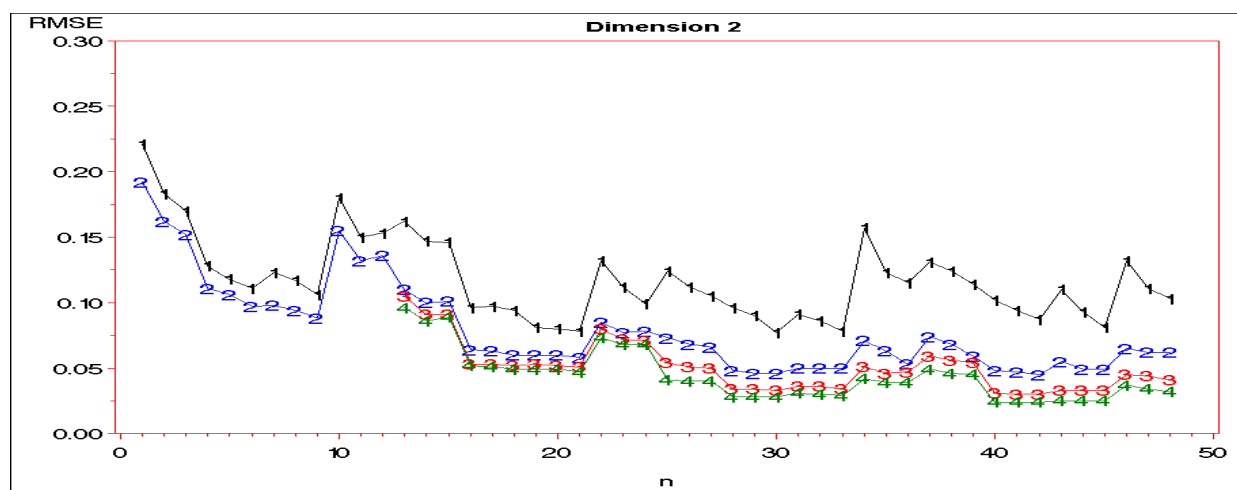
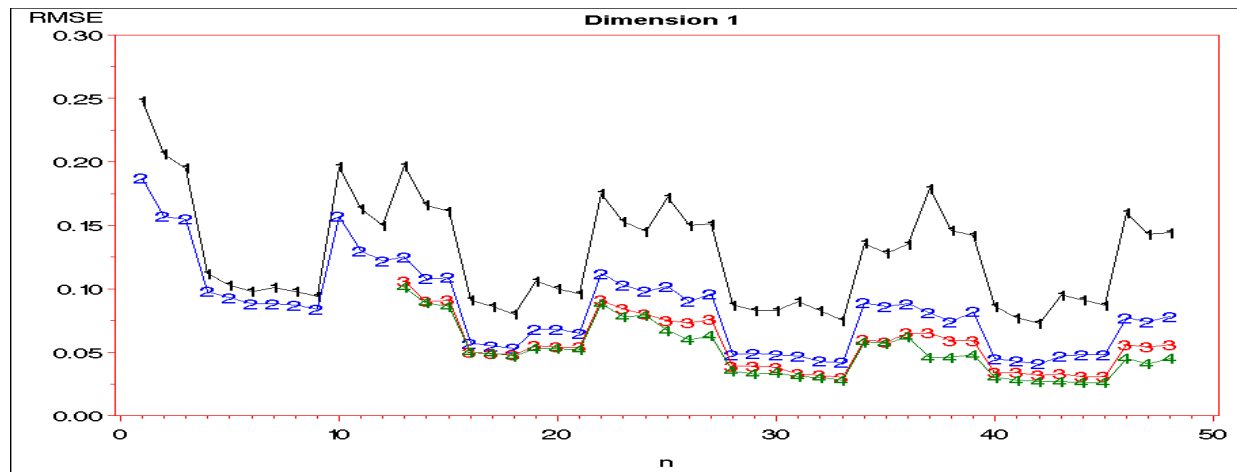


Figure 4.2.Root Mean Squared Error under different test length
(line1: length =5, line2: length =10, line3: length =15, line4: length =20)

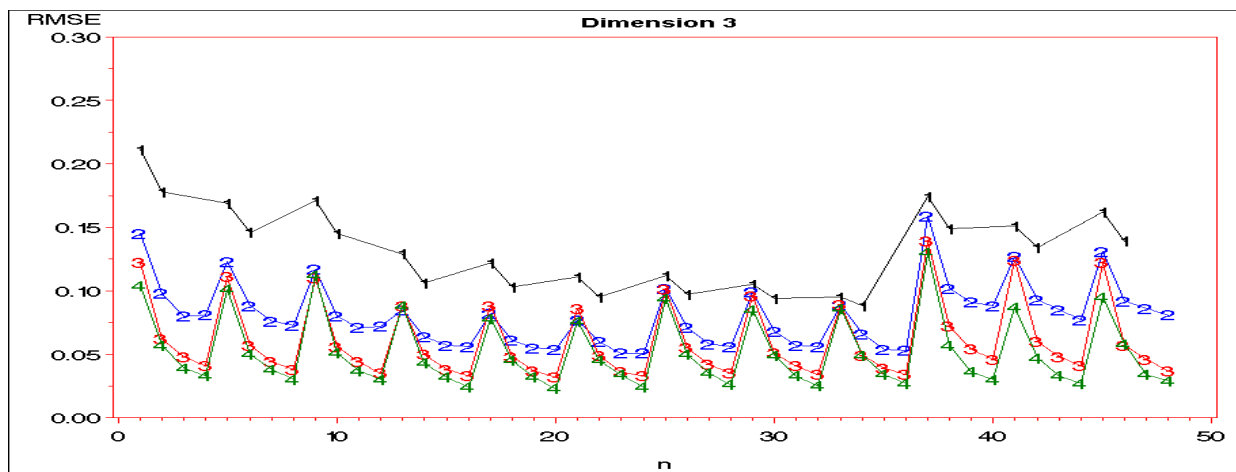
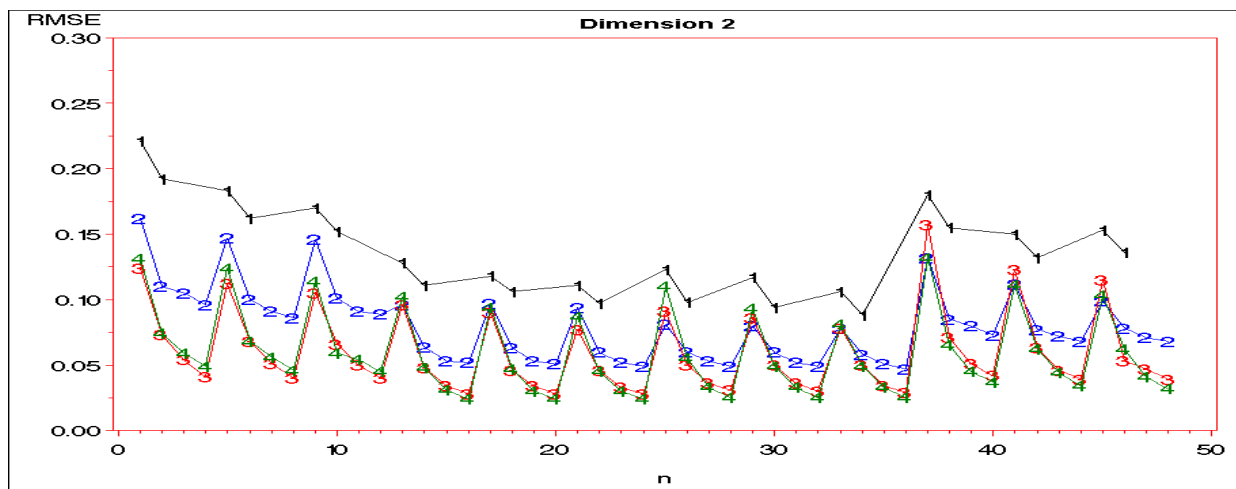
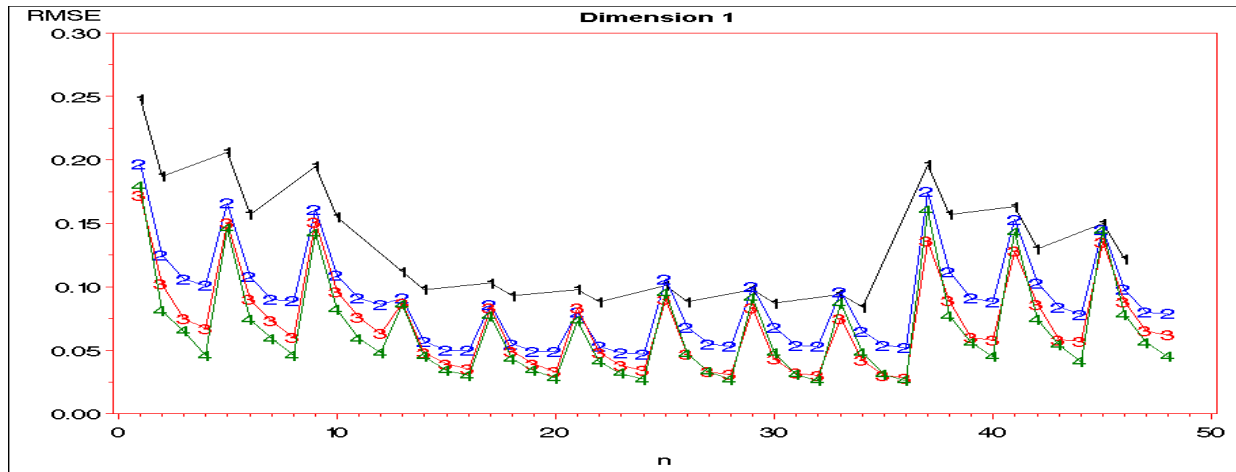


Figure 4.3.Root Mean Squared Error under different item pool size
(line1: pool =10, line2: pool =20, line3: pool =50, line4: pool =100)

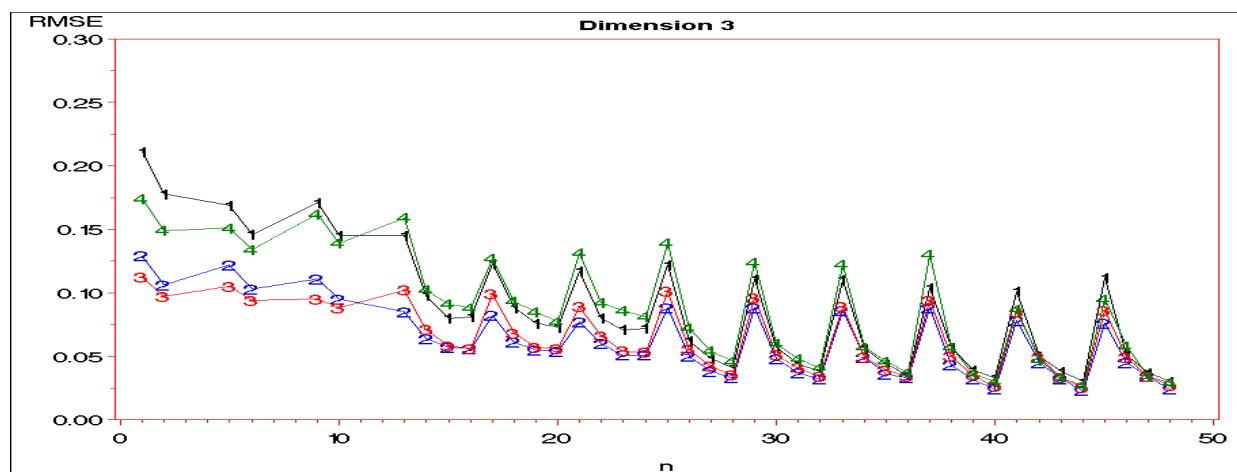
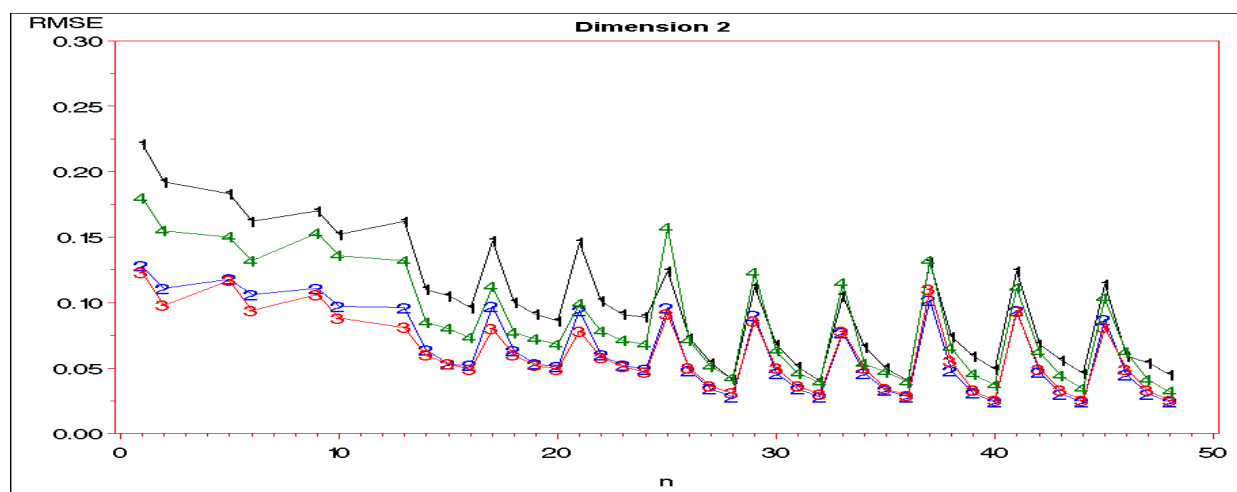
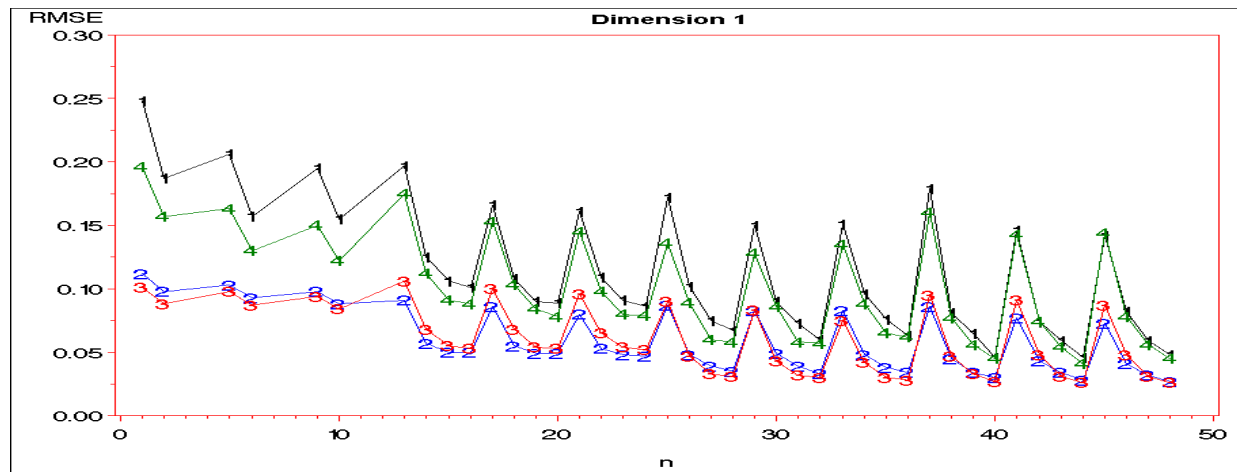


Figure 4.4. Root Mean Squared Error under different ability level
(line1: ability =1, line2: ability =2, line3: ability =3, line4: ability =4)

As expected, the correlation between dimensions did have an impact on the RMSE. By comparing the RMSE values in Table 4.1 under different correlations for fixed item pool sizes, test lengths, and ability levels, several trends can be observed:

- 1) RMSE decreased as the correlations between dimensions increased, but more change was observed when the correlation changed from 0.0 to 0.4 than from 0.4 to 0.7. In addition, the impact was not notable for extreme ability levels (ability groups 1 and 4), for short test lengths (5 items), and small item pool sizes (10 and 20 items).
- 2) As test length increased to 15 items, RMSE decreased. There was little change when increasing the test from 15 to 20 items.
- 3) As item pool size increased to 50 items, there was a decrease in RMSE.
- 4) As would be expected, RMSE was smaller in the middle range ability levels (groups 2 and 3).

Figures 4.1 to 4.4 support these findings as well. As shown in Figure 4.1, line 1 (correlations = 0.0) had a larger RMSE than line 2 (correlations = 0.4) and line 3 (correlations = 0.7) for all three dimensions. This was more evident when the item pool size was small (10 items) as shown in Figure 4.1 where $n < 17$. When the item pool size was large (50 items or 100 items – $n > 32$), the differences in RMSE between different correlations were negligible. This is reasonable since there are numerous “good” items (items that match ability estimates) to be selected when the item pool size was “large”. The decrease in RMSE was more significant when correlations between dimensions changed from 0.0 to 0.4 than that observed when correlations between dimensions changed from 0.4 to 0.7. In Figure 4.1, the vertical distance between line 1 (correlations = 0.0) and line 2 (correlations = 0.4) was larger than the distance between line 2 (correlations = 0.4) and line 3 (correlations = 0.7). This was particularly the case, when ability level was in the extreme range ($\theta < -1$ or $\theta > 1$), or where n was in the ranges: 1-4, 12-20, 28-36, 44-52, and 61-64.

Figure 4.2 illustrates the effect of longer tests on the accuracy of ability estimates. A comparison of the RMSE based on different test lengths when item pool size, correlations between dimensions, and ability level were fixed, revealed that the RMSE tended to decrease when test length increased for all three dimensions. However, the decrease in the RMSE was more note worthy when test length was short (5 items or 10 items) as opposed to when test

length was longer (15 items or 20 items). As indicated in Figure 4.2, the vertical distances between line 1 (test length = 5) and line 2 (test length = 10) was much larger than the distance between line 3 (test length = 15) and line 4 (test length = 20). In addition, the decrease in the RMSE between different test lengths was more evident when the item pool size was large (50 items or 100 items – $n > 24$), than when the item pool size was small (10 items or 20 items – $n < 25$).

Figure 4.3 illustrates the effect of item pool size on RMSE. When comparing line 1 (item pool size = 10), line 2 (item pool size = 20), line 3 (item pool size = 50), and line 4 (item pool size = 100), it was clear that the RMSE tended to decrease when item pool size increased for all three dimensions. The decrease in the RMSE was more significant when the item pool size was small (item pool size = 10 or 20) than when the item pool size was large (item pool size = 50 or 100). For example, the vertical distance between line 1 and line 2 was greater than the distance between line 2 and line 3, and the distance between line 3 and line 4 was the smallest. In addition, the decrease in the RMSE between different item pool sizes was more significant for examinees who had ability levels in extreme ranges ($\theta < -1$ or $\theta > 1$ - $n < 13$ and $n > 36$), than examinees with ability levels in the middle range ($-1 < \theta < 1$ - $12 < n < 36$).

Finally, Figure 4.4 illustrates the effect of ability level on the RMSE. As can be seen, larger RMSE values were observed for examinees with extreme ability (line 1 and line 4) than for examinees with ability in the middle range (line 2 and line 3). Further, the differences in RMSE between different ability levels was more evident when the item pool size was smaller (10 items – $n < 17$) than when the item pool size was larger (>10 items – $n > 16$). In addition, as the number of estimated dimensions increased, RMSE decreased. When comparing the three panels in Figure 4.4, RMSE differences decreased as the number of dimensions increased from 1 to 3. This finding is specific to the design of this study since in this study only the item responses from the first dimension was factored into the estimation of ability for dimension 1. For the second dimension, however, the item responses from the dimension and the relationship between the first and second dimension were factored into the estimation of ability. Finally, for the third dimension, the relationships between all three dimensions were factored into ability estimation.

The difference in the RMSE values for the different ability groups may be explained by examining the test information function. Figure 4.5 presents the test information functions for the 10 items used in this study. As can be seen, more information for estimating ability was available

in the middle range ($-1 < \theta < 1$) than in the extreme range ($\theta < -1$ and $\theta > 1$). This is typical of testing applications and would result in more accurate ability estimates and thus smaller RMSE values in the middle range of ability. However, there were also some small differences between the two extreme ability groups ($\theta < -1$ and $\theta > 1$), with RMSE values tending to be smaller for $\theta > 1$. This can also be explained by the test information function. From Figure 4.5, slightly more information was available at the higher ability range than the lower ability range.

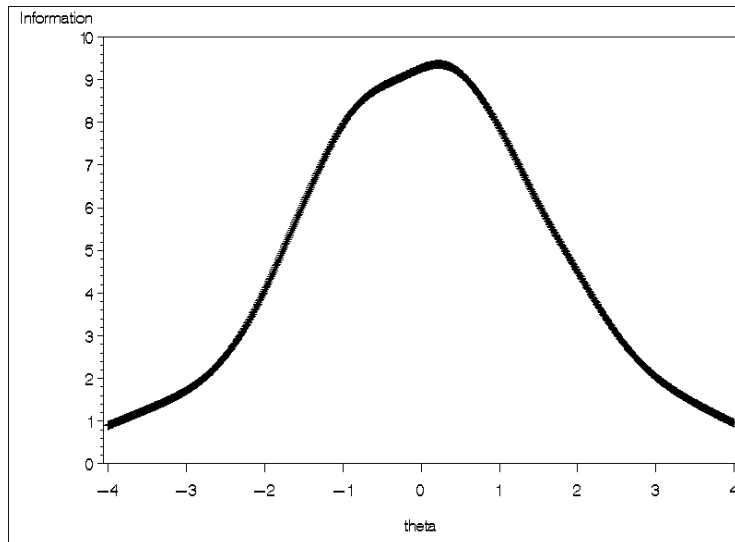


Figure 4.5. Test Information for 10 items

A mixed design analysis of variance (ANOVA) was conducted in order to evaluate the significance of the factors under study. Item pool size and correlation between dimensions were between-subject factors. The number of estimated dimensions and test lengths were within-subject factors. It should be noted that the squared error for each examinee served as the outcome measure for the ANOVA ($Squared\ Error = (\hat{\theta}_{ij} - \theta_{ij})^2$, where $\hat{\theta}_{ij}$ is the estimated ability value and θ_{ij} is the true ability value). RMSE could not be used directly in the ANOVA since it reflected an outcome measure aggregated across examinees. Also, ability level was excluded from the analysis because examinees' ability level group was not fixed for each dimension.

Examinees could theoretically be assigned to different ability groups for each of the three dimensions. As discussed above, the study included combinations of conditions that were missing by design. Thus, the degrees of freedom were adjusted and type IV sum of squares were used. Finally, the assumption of sphericity was not satisfied. Therefore, Greenhouse-Geisser results were reported.

Table 4.3 summarizes the results from ANOVA. Using $\alpha = .05$, the following effects were statistically significant:

- One three-way interaction effect: test length * number of dimensions for estimation* item pool size,
- Four two-way interaction effects: test length * item pool size, test length * correlations between dimensions, number of dimensions for estimation * correlations between dimensions, and test length * correlations between dimensions,
- Four main effects: test length, number of dimensions for estimation, item pool size, and correlations between dimensions.

R^2 and effect size (η^2) are also listed in Tale 4.4. R^2 is the relative predictive power of a model. It is defined by

$$R^2 = 1 - \frac{SS_{error}}{SS_{total}} \quad (4.1)$$

Effect size can provide additional information on the significance of effects. The proportion of the total variance that is attributed to an effect is denoted by η^2 . It is calculated as the ratio of the effect variance (SS_{effect}) to the total variance (SS_{total}), or

$$\eta^2 = SS_{effect} / SS_{total} \quad (4.2)$$

Using η^2 as the measure of effect size, the effect of test length accounted for most of the variance (6%).

Note that while some significant interaction terms were noted, the effect sizes associated with these effects were essentially zero. This indicated an effect that may have no practical interpretative value. The finding of significant effects with low effect sizes is not unusual in simulation studies (Harwell, Stone, Hsu, & Kirisci, 1996).

Table 4.3. ANOVA Results

Source	df	F	P	η^2
Length	1.402	3870.48	0.00	.06
Length * Pool	2.803	51.95	0.00	0
Length * Correlation	2.803	15.22	0.00	0
Length * Pool * Correlation	5.606	.47	.820	0
Order	1.979	10.75	0.00	0
Order * Pool	3.958	2.43	.046	0
Order * Correlation	3.958	0.12	.975	0
Order * Pool * Correlation	7.916	0.21	.990	0
Length * Order	2.789	3.91	.010	0
Length * Order * Pool	5.578	3.32	.004	0
Length * Order * Correlation	5.578	.37	.887	0
Length * Order * Pool * Correlation	11.155	.40	.960	0
Pool	2	143.67	0.00	.01
Correlation	2	11.22	0.00	0
Pool * Correlation	4	.19	.946	0
$R^2 = 0.3$				

4.1.2 Bias Measure

Defined by Equation 3.47, bias was calculated as the mean difference between estimated ability and true ability for each combination (3 correlations between dimensions x 4 test length x 4 item pool size x 4 ability level x 3 dimensions). These results are presented in Table 4.4. Figures 4.6 through Figure 4.9 also illustrate the bias for each combination under different correlations, test length, item pool size, ability level, and dimensions. The “n” on the horizontal axes identify the experimental conditions and may be translated using Table 4.2.

As shown in Table 4.4, ability group had a large impact on bias. By comparing the bias values in Table 4.4 under different ability group for fixed correlations between dimensions, item pool sizes, and test lengths, several trends can be observed:

- 1) When ability was negative (ability = 1 or 2), bias values were positive (which means $\hat{\theta} > \theta$). When ability level was positive (ability = 3 or 4), bias values were

negative (which means $\hat{\theta} < \theta$). This is consistent with Bayesian estimation methods which “shrink” estimates toward the mean of the prior distribution.

- 2) The absolute values of bias were smaller for examinees with ability values in the middle range ($-1 < \theta < 1$) than examinees with ability values in the extreme range ($\theta < -1$ or $1 < \theta$). This would also be expected given the information function discussed above.
- 3) The absolute value of bias decreased as the test length increased, but more change was observed when the test length changed from 5 items to 10 items. The impact was not notable for examinees with extreme ability levels (ability = 1 or 4).
- 4) The effect due to test length may be observed but this effect is less for longer tests (15 or 20 items) and less for examinees in the middle range of ability.
- 5) A modest effect due to the correlation between dimensions was found for tests where the correlation increased from 0 to .4. The effect was mitigated as test length increased.
- 6) The number of estimated dimensions also appeared to influence bias. This effect was particularly notable for shorter tests and in the more extreme ability groups.

Figures 4.6 to 4.9 support these findings as well. As shown in Figure 4.6, line 1 (ability = 1, $\theta < -1$) and line 2 (ability = 2, $-1 < \theta < 0$) were above 0 and line 3 (ability = 3, $0 < \theta < 1$) and line 4 (ability = 4, $1 < \theta$) were below 0. This indicates that the ability was over-estimated for examinees with negative true ability value and the ability was under-estimated for examinees with positive true ability value. In addition, as shown in Figure 4.6, line 2 (ability = 2, $-1 < \theta < 0$) and line 3 (ability = 3, $0 < \theta < 1$) had smaller absolute bias values than line 1 (ability = 1, $\theta < -1$) and line 4 (ability = 4, $1 < \theta$), which indicated that ability was estimated more accurately for examinees with ability values in the middle range ($-1 < \theta < 1$) than examinees with ability values in the extreme range ($\theta < -1$ or $1 < \theta$). This is consistent with the RMSE results. It can also be observed that the differences in bias between different ability levels were larger when comparing a small item pool size (10 items - $n < 13$) with large item pool size (>10 items - $n > 12$).

Figure 4.7 illustrates the effect of longer tests on the accuracy of ability estimates. By comparing line 1 (test length = 5 items), line 2 (test length = 10 items), line 3 (test length = 15 items), and line 4 (test length = 20 items), when item pool size, correlations between dimensions,

and ability level were fixed, it can be observed that the absolute value of bias tended to decrease when test length increased. The decrease in the bias was noteworthy when test length was short (5 items or 10 items) as opposed to when test length was long (15 items or 20 items). In addition, the decrease in bias was more pronounced when ability level was in the extreme range ($\theta < -1$ or $1 < \theta - n$ in the ranges: 1-3, 10-16, 22-27, 34-39, and 46-48). When the ability level was in the medium range ($-1 < \theta < 1 - n$ in other ranges besides those noted above), similar bias values were observed. This also was consistent with the RMSE results. In addition, as for the RMSE measure, the bias measure decreased as the number of estimated dimensions increased. As shown in Figure 4.7, for most cases, the absolute values of bias on dimension 1 were the largest, followed by dimension 2, and then followed by dimension 3.

Unlike the RMSE results, the absolute values for bias were not consistently related to item pool size. As shown in Figure 4.8, the absolute values of bias for the item pool size = 10 (line 1) were larger than that for other item pool sizes for some cases. However, the absolute values of bias for item pool size = 50 (line 3) was not consistently larger than that for item pool size = 100 (line 4). This could be due to positive bias and negative bias canceling each other out.

Also, the impact of the correlations between dimensions on the bias measure was not as notable as found for the RMSE measure. As indicated in Figure 4.9, the difference in bias between different correlations was negligible for most cases. A difference in bias values between different correlations was observable only when the item pool size was small (10 items or 20 items - $n < 33$) and ability level was in the extreme range ($\theta < -1$ or $1 < \theta - n$ in the ranges: 1-4, 13-20, and 29-32).

Note that an ANOVA was not conducted for the bias measure. When collapsing across ability groups as done for the RMSE measure, no effects for these factors were observed. This can be seen by examining the “Total” rows in Table 4.4.

Table 4.4. Bias

Pool Size	Ability Level	Correlation	Dimension											
			1				2				3			
			Test Length				Test Length				Test Length			
			5	10	15	20	5	10	15	20	5	10	15	20
10	1	0	0.28	0.21			0.27	0.24			0.23	0.22		
		0.4	0.23	0.16			0.21	0.18			0.17	0.17		
		0.7	0.24	0.18			0.22	0.19			0.21	0.19		
	2	0	0.11	0.07			0.06	0.04			0.05	0.04		
		0.4	0.10	0.06			0.05	0.03			0.05	0.04		
		0.7	0.09	0.07			0.06	0.04			0.06	0.05		
	3	0	-0.06	-0.03			-0.08	-0.05			-0.01	0		
		0.4	-0.05	-0.02			-0.07	-0.04			0	0		
		0.7	-0.05	-0.03			-0.07	-0.05			-0.02	-0.01		
	4	0	-0.28	-0.22			-0.26	-0.23			-0.22	-0.21		
		0.4	-0.22	-0.17			-0.20	-0.19			-0.16	-0.16		
		0.7	-0.22	-0.18			-0.22	-0.20			-0.19	-0.17		
	Total	0	0.01	0.01			-0.01	0			0.02	0.02		
		0.4	0.01	0.01			-0.01	-0.01			0.02	0.02		
		0.7	0.01	0.01			-0.01	-0.01			0.02	0.02		
20	1	0	0.28	0.22	0.17	0.15	0.16	0.14	0.15	0.14	0.19	0.11	0.09	0.09
		0.4	0.24	0.18	0.14	0.12	0.12	0.11	0.13	0.12	0.14	0.08	0.06	0.07
		0.7	0.25	0.19	0.15	0.14	0.15	0.14	0.14	0.14	0.14	0.08	0.09	0.09
	2	0	0.07	0.06	0.02	0.01	0.02	0.01	0.01	0	0.06	0.03	0.02	0.02
		0.4	0.06	0.05	0.01	0.01	0.01	0	0	0	0.06	0.03	0.02	0.02
		0.7	0.06	0.04	0.01	0.01	0.02	0.01	0.01	0.01	0.05	0.02	0.02	0.02
	3	0	-0.09	-0.09	-0.05	-0.04	-0.02	-0.01	-0.01	-0.01	-0.01	0	-0.01	-0.01
		0.4	-0.09	-0.07	-0.04	-0.03	-0.01	-0.01	-0.01	0	0	-0.01	-0.01	-0.01
		0.7	-0.08	-0.07	-0.04	-0.04	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02
	4	0	-0.22	-0.17	-0.12	-0.11	-0.18	-0.13	-0.12	-0.11	-0.21	-0.14	-0.13	-0.13
		0.4	-0.18	-0.14	-0.10	-0.08	-0.12	-0.08	-0.09	-0.08	-0.15	-0.11	-0.10	-0.10
		0.7	-0.19	-0.15	-0.11	-0.1	-0.11	-0.10	-0.10	-0.09	-0.17	-0.12	-0.13	-0.12
	Total	0	0	0	-0.01	-0.01	-0.01	0	0	0	0.01	0	-0.01	-0.01
		0.4	-0	0	-0.01	-0.01	0	0	0	0	0.01	0	-0.01	-0.01
		0.7	0	0	-0.01	-0.01	0.01	0	0.01	0.01	0.01	0	-0.01	-0.01

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,
ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.
Total means all ability levels.

Table 4.4. Continued

Pool Size	Ability Level	Correlation	Dimension											
			1				2				3			
			Test Length				Test Length				Test Length			
			5	10	15	20	5	10	15	20	5	10	15	20
50	1	0	0.21	0.16	0.12	0.12	0.12	0.06	0.05	0.03	0.16	0.09	0.08	0.07
		0.4	0.19	0.14	0.12	0.11	0.10	0.05	0.04	0.02	0.15	0.08	0.06	0.05
		0.7	0.22	0.17	0.13	0.13	0.15	0.06	0.05	0.03	0.17	0.09	0.07	0.05
	2	0	0.06	0.06	0.06	0.05	0.05	0.03	0.02	0.01	0.04	0.04	0.03	0.04
		0.4	0.05	0.05	0.06	0.06	0.03	0.02	0.02	0.01	0.02	0.03	0.03	0.03
		0.7	0.05	0.05	0.06	0.06	0.04	0.02	0.02	0.01	0.04	0.03	0.03	0.02
	3	0	-0.05	-0.05	-0.04	-0.05	-0.04	-0.02	-0.01	-0.01	-0.06	-0.04	-0.02	-0.03
		0.4	-0.05	-0.05	-0.05	-0.05	-0.03	-0.02	-0.01	0	-0.05	-0.02	-0.02	-0.02
		0.7	-0.05	-0.05	-0.05	-0.06	-0.02	-0.01	-0.01	0	-0.05	-0.02	-0.02	-0.01
	4	0	-0.18	-0.15	-0.13	-0.12	-0.20	-0.09	-0.06	-0.04	-0.15	-0.07	-0.06	-0.05
		0.4	-0.15	-0.13	-0.12	-0.11	-0.13	-0.06	-0.04	-0.03	-0.10	-0.06	-0.05	-0.04
		0.7	-0.18	-0.15	-0.14	-0.14	-0.17	-0.07	-0.05	-0.04	-0.14	-0.07	-0.06	-0.03
	Total	0	0	0	0	0	-0.01	0	0	0	0	0.01	0.01	0.01
		0.4	0	0	0	0	0	0	0	0	0	0.01	0.01	0.01
		0.7	0	0	0	0	0.01	0	0	0	0	0.01	0.01	0.01
100	1	0	0.25	0.14	0.11	0.08	0.18	0.11	0.08	0.06	0.16	0.08	0.06	0.06
		0.4	0.21	0.13	0.1	0.08	0.16	0.08	0.06	0.05	0.13	0.06	0.03	0.04
		0.7	0.21	0.14	0.12	0.09	0.15	0.09	0.07	0.05	0.16	0.07	0.06	0.04
	2	0	0.07	0.06	0.05	0.05	0.03	0.02	0.01	0.01	0.02	0.01	0.01	0
		0.4	0.06	0.06	0.05	0.04	0.04	0.01	0.01	0.01	0.02	0.01	0	-0.01
		0.7	0.05	0.06	0.05	0.05	0.05	0.02	0.02	0.01	0.01	0.02	0.01	-0.01
	3	0	-0.05	-0.05	-0.05	-0.04	-0.05	-0.02	-0.02	-0.02	-0.06	-0.03	-0.02	-0.01
		0.4	-0.05	-0.04	-0.05	-0.04	-0.05	-0.02	-0.02	-0.01	-0.04	-0.01	-0.01	-0.01
		0.7	-0.05	-0.04	-0.05	-0.05	-0.05	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01
	4	0	-0.22	-0.15	-0.12	-0.09	-0.13	-0.07	-0.05	-0.05	-0.19	-0.09	-0.06	-0.05
		0.4	-0.19	-0.13	-0.11	-0.09	-0.12	-0.05	-0.05	-0.04	-0.08	-0.05	-0.04	-0.04
		0.7	-0.20	-0.13	-0.12	-0.11	-0.14	-0.08	-0.05	-0.05	-0.16	-0.07	-0.06	-0.05
	Total	0	0.01	0	0	0	0	0	0	0	-0.02	-0.01	0	0
		0.4	0.01	0	0	0	0	0	0	0	0	0	0	-0.01
		0.7	0	0.01	0	0	0	0	0	0	0	0	0	-0.01

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,
ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.
Total means all ability levels.

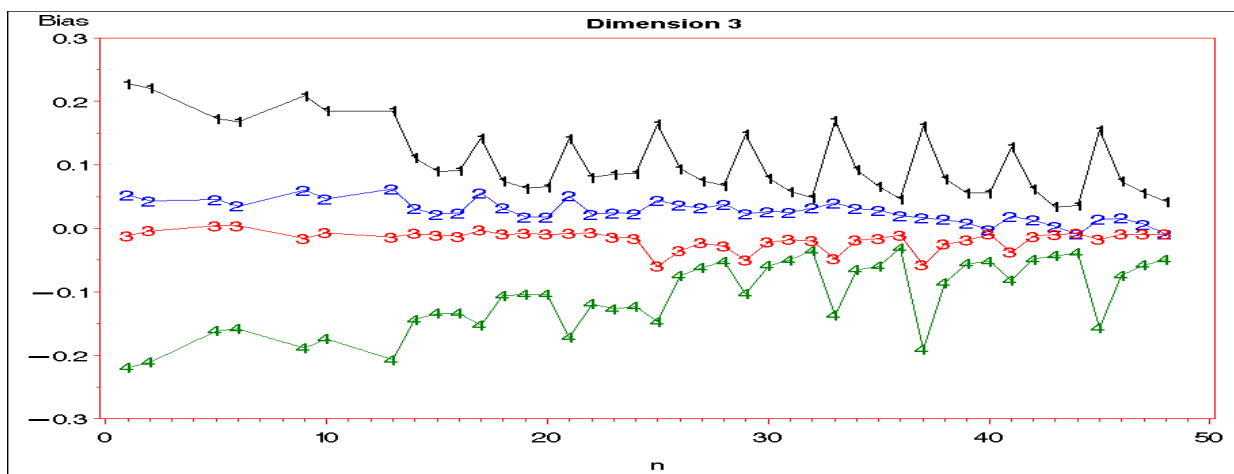
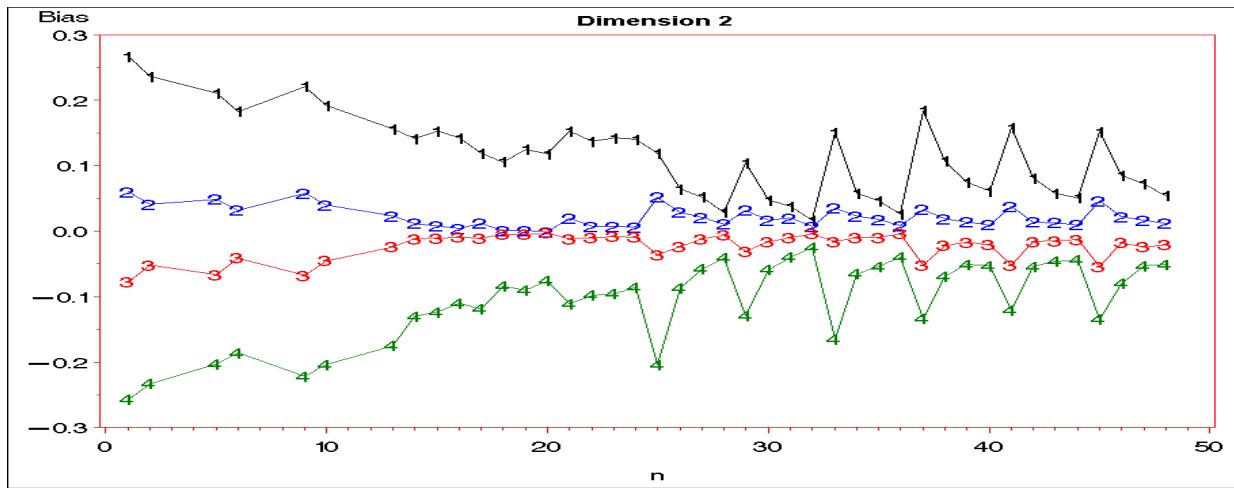
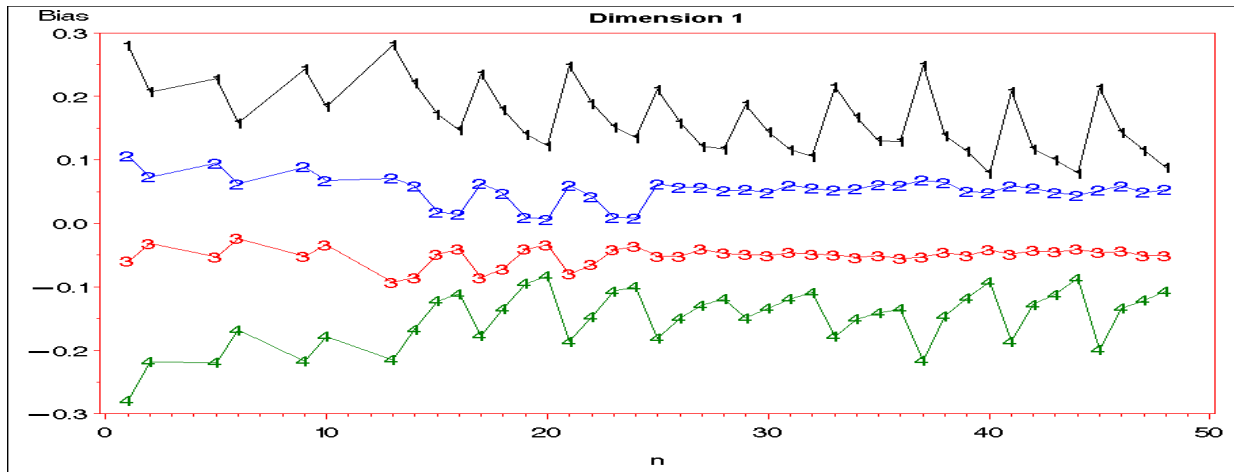


Figure 4.6. Bias under different ability level
(line1: ability =1, line2: ability =2, line3: ability =3, line4: ability =4)

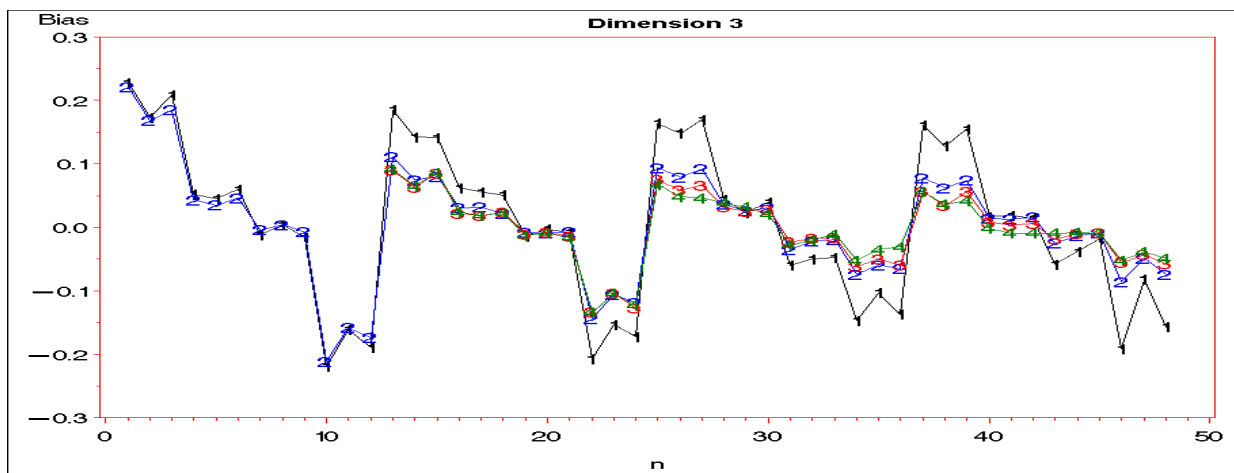
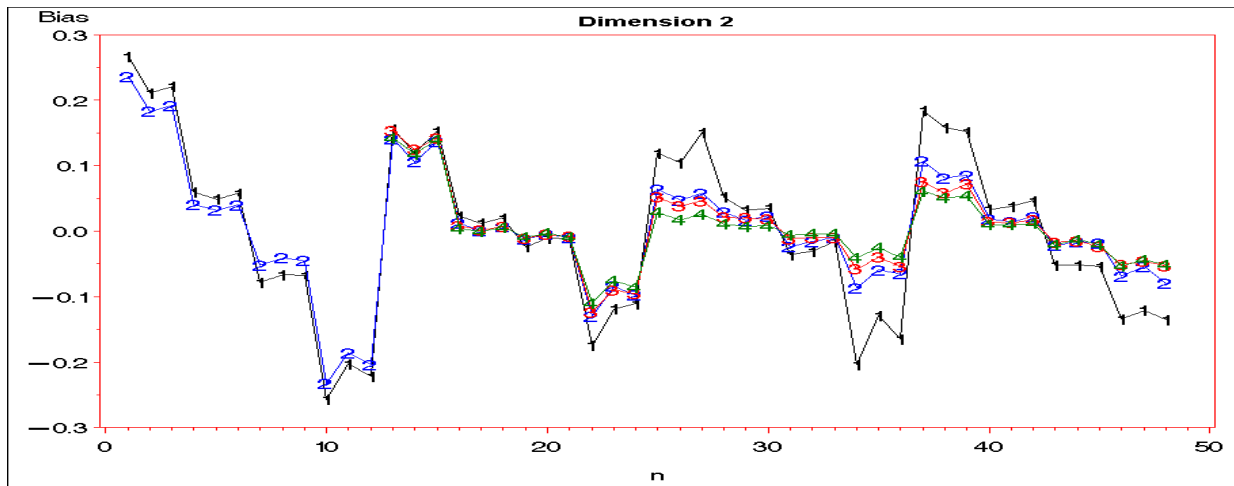
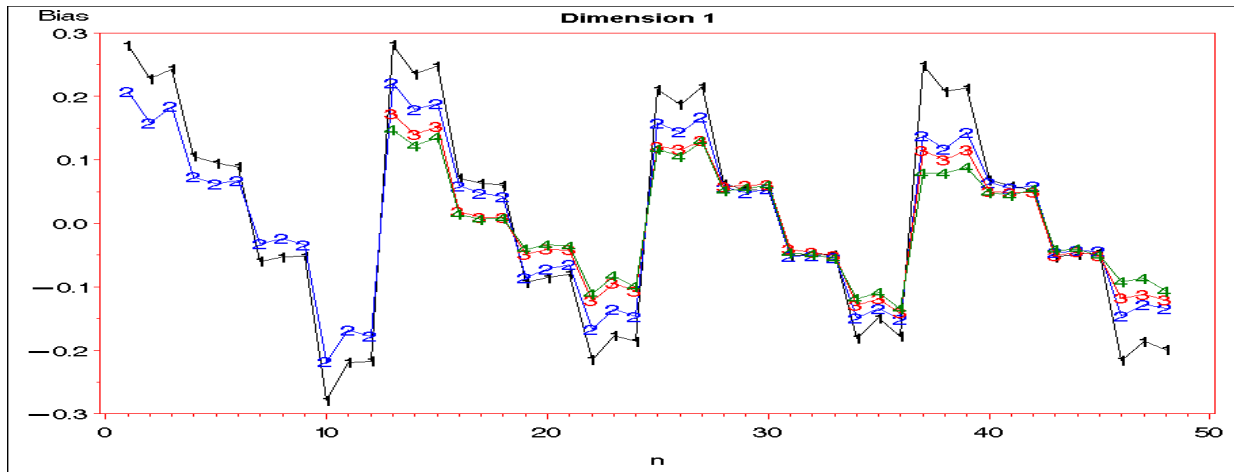


Figure 4.7. Bias under different test length
 (line1: length =5, line2: length =10, line3: length =15, line4: length =20)

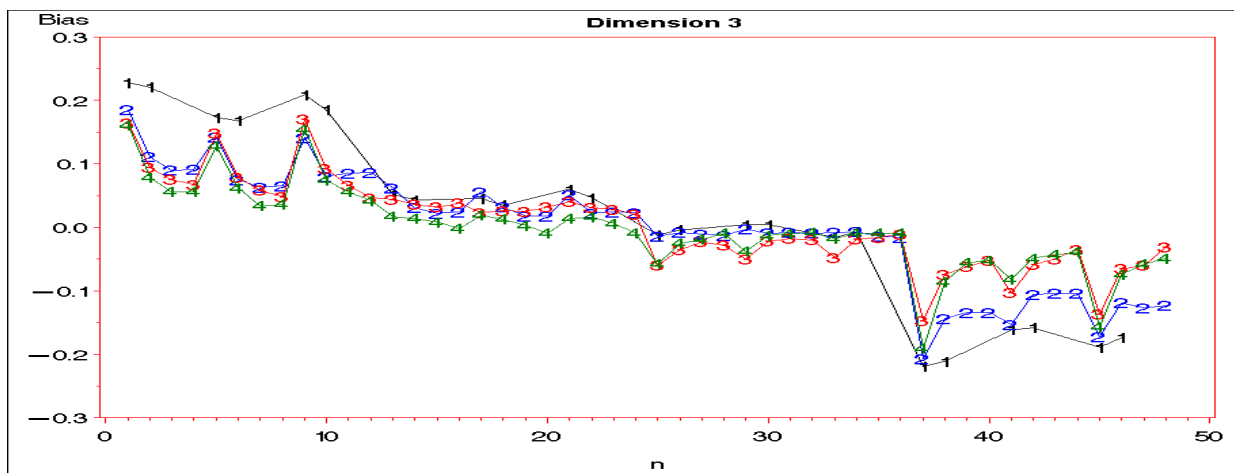
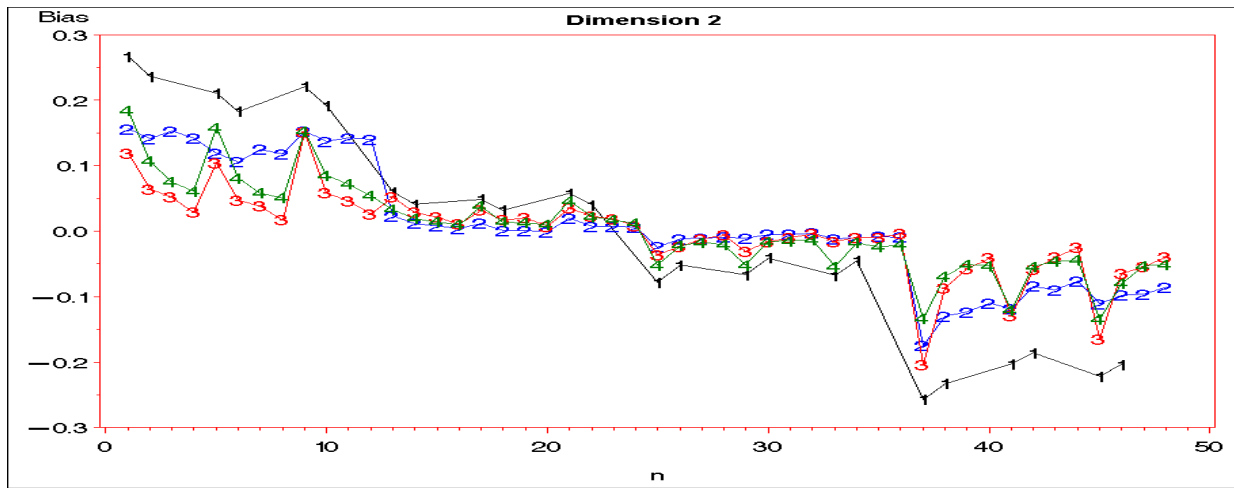
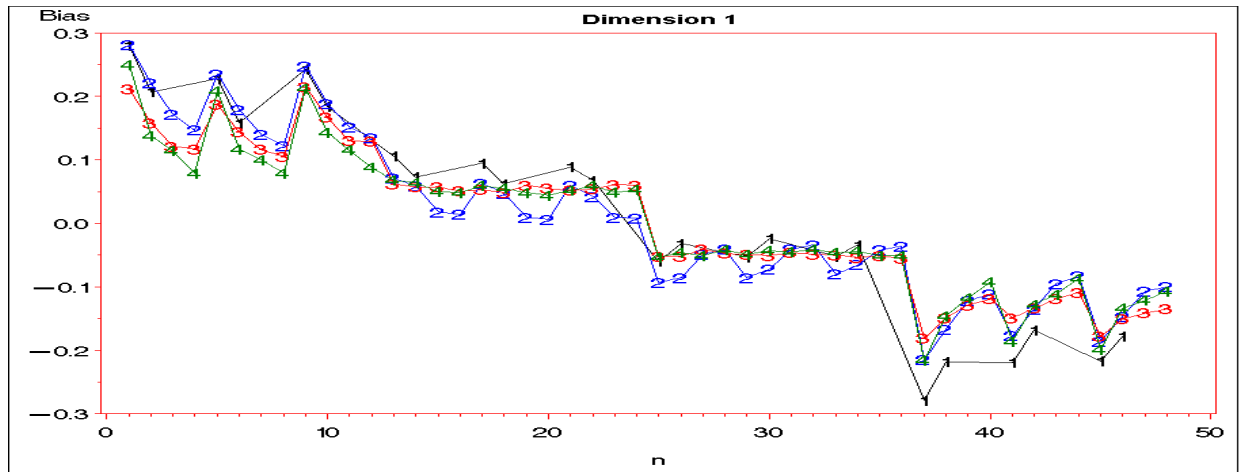


Figure 4.8. Bias under different item pool size
(line1: pool =10, line2: pool =20, line3: pool =50, line4: pool =100)

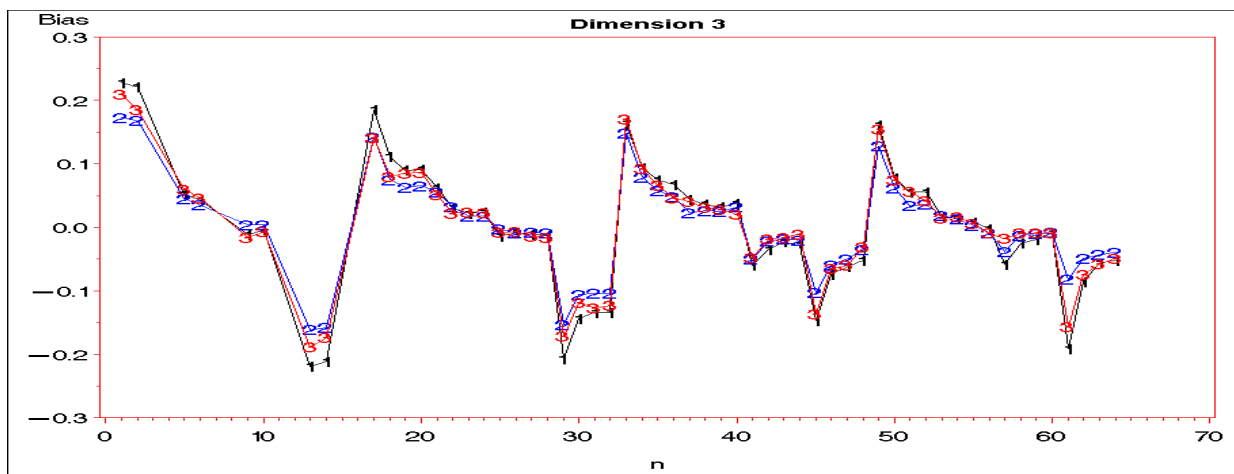
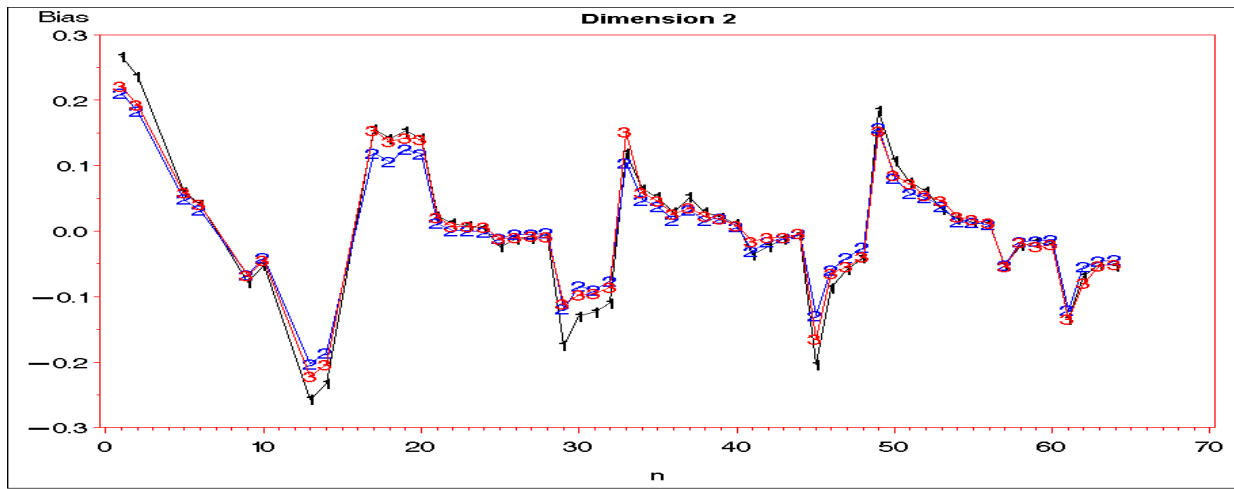
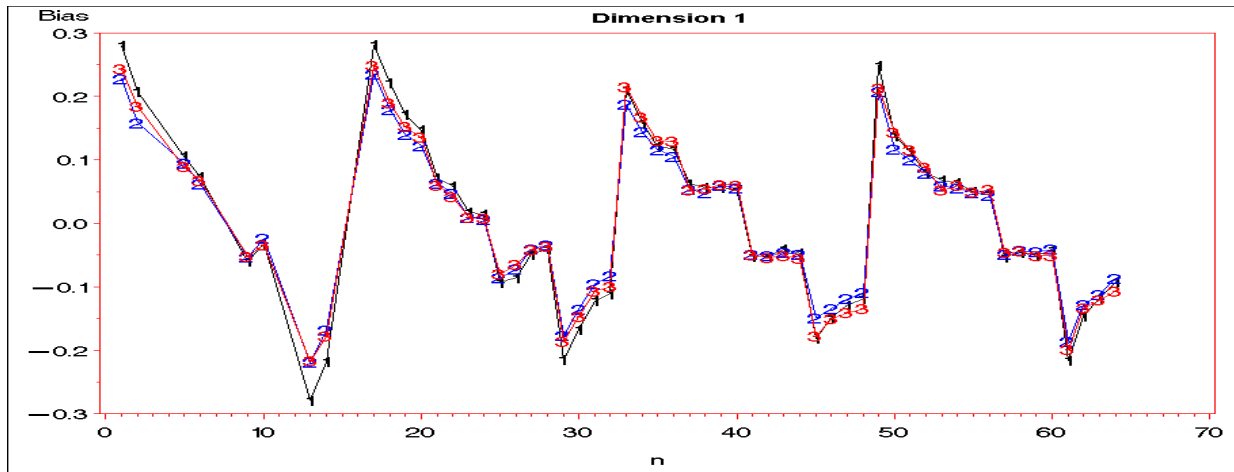


Figure 4.9. Bias under different correlation
(line1: correlation = 0.0, line2: correlation = 0.4, line3: correlation = 0.7)

4.1.3 Pearson Correlation Measure

For each combination of item pool size, test length, and correlations between dimensions, 1000 vectors of true ability values were simulated and MCAT was used to estimate the examinee's abilities on three dimensions. The Pearson correlation between true ability and estimated ability was calculated for each combination. The results are presented in Table 4.5.

In Table 4.5, the Pearson correlations are summarized according to different ability groups. By comparing the values between different ability groups, larger Pearson correlations were found for examinees with extreme ability (ability = 1 or 4) than for examinees with ability in the middle range (ability = 2 or 3). When compared with the results for all examinees (ability = total), the Pearson correlations for all examinees were larger. This may be explained by the "restriction of range" factor which is known to affect the correlation statistic. Figures 4.10, 4.11, and 4.12 illustrate this effect by plotting true ability on dimension 1 and estimated ability on dimension 1 when test length was 5 items. Under no restriction in the ability range (see Figure 4.1), there was a strong relationship across the range from -4 to 4. However, a narrower range in ability (from -1 to -4 for the extreme group, see Figure 4.11) attenuates the relationship, and an even narrower range in ability (0 to -1 for a middle group, see Figure 4.12) even further attenuates the relationship.

Table 4.5. Pearson Correlation

Pool Size	Ability Level	Correlation	Dimension											
			1				2				3			
			Test Length				Test Length				Test Length			
			5	10	15	20	5	10	15	20	5	10	15	20
10	1	0	0.61	0.66			0.63	0.65			0.65	0.71		
		0.4	0.67	0.70			0.67	0.69			0.70	0.74		
		0.7	0.70	0.73			0.71	0.73			0.72	0.75		
	2	0	0.59	0.63			0.59	0.63			0.61	0.64		
		0.4	0.62	0.65			0.61	0.64			0.62	0.66		
		0.7	0.62	0.66			0.62	0.65			0.62	0.67		
	3	0	0.64	0.67			0.64	0.68			0.59	0.60		
		0.4	0.65	0.69			0.66	0.70			0.61	0.61		
		0.7	0.66	0.69			0.68	0.71			0.61	0.62		
	4	0	0.69	0.70			0.61	0.65			0.70	0.74		
		0.4	0.71	0.73			0.66	0.68			0.71	0.74		
		0.7	0.73	0.75			0.67	0.69			0.70	0.73		
	Total	0	0.93	0.94			0.93	0.94			0.93	0.94		
		0.4	0.94	0.95			0.93	0.94			0.94	0.95		
		0.7	0.94	0.95			0.94	0.95			0.94	0.95		
20	1	0	0.77	0.84	0.84	0.83	0.73	0.81	0.82	0.83	0.73	0.79	0.82	0.81
		0.4	0.80	0.85	0.85	0.85	0.75	0.82	0.84	0.84	0.75	0.80	0.82	0.83
		0.7	0.81	0.85	0.85	0.86	0.75	0.82	0.84	0.85	0.76	0.81	0.83	0.83
	2	0	0.63	0.73	0.77	0.77	0.61	0.70	0.74	0.75	0.67	0.73	0.75	0.76
		0.4	0.65	0.74	0.78	0.78	0.62	0.71	0.74	0.75	0.69	0.75	0.76	0.77
		0.7	0.66	0.75	0.78	0.78	0.62	0.71	0.75	0.76	0.71	0.75	0.77	0.78
	3	0	0.64	0.71	0.75	0.76	0.68	0.77	0.78	0.79	0.61	0.70	0.74	0.74
		0.4	0.65	0.71	0.76	0.76	0.69	0.77	0.79	0.80	0.63	0.72	0.75	0.75
		0.7	0.65	0.71	0.75	0.76	0.70	0.78	0.79	0.80	0.65	0.73	0.76	0.75
	4	0	0.65	0.78	0.79	0.79	0.69	0.80	0.80	0.80	0.72	0.80	0.81	0.82
		0.4	0.68	0.78	0.80	0.81	0.74	0.80	0.81	0.81	0.76	0.81	0.82	0.84
		0.7	0.70	0.79	0.81	0.81	0.74	0.79	0.81	0.81	0.76	0.80	0.82	0.83
	Total	0	0.93	0.96	0.97	0.97	0.95	0.96	0.97	0.97	0.94	0.96	0.97	0.97
		0.4	0.94	0.96	0.97	0.97	0.95	0.96	0.97	0.97	0.95	0.96	0.97	0.97
		0.7	0.94	0.96	0.97	0.97	0.95	0.97	0.97	0.97	0.95	0.96	0.97	0.97

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,
ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.
Total means all ability levels.

Table 4.5. Continued

Pool Size	Ability Level	Correlation	Dimension											
			1				2				3			
			Test Length				Test Length				Test Length			
			5	10	15	20	5	10	15	20	5	10	15	20
50	1	0	0.75	0.83	0.87	0.88	0.65	0.78	0.83	0.87	0.77	0.88	0.90	0.91
		0.4	0.78	0.85	0.87	0.90	0.68	0.79	0.84	0.87	0.79	0.88	0.91	0.92
		0.7	0.78	0.85	0.87	0.90	0.73	0.80	0.85	0.87	0.81	0.88	0.91	0.92
	2	0	0.66	0.78	0.82	0.84	0.61	0.77	0.82	0.85	0.67	0.79	0.83	0.85
		0.4	0.67	0.77	0.82	0.84	0.63	0.78	0.83	0.85	0.69	0.79	0.84	0.85
		0.7	0.66	0.77	0.82	0.84	0.66	0.78	0.83	0.85	0.68	0.79	0.84	0.85
	3	0	0.62	0.75	0.82	0.84	0.69	0.78	0.83	0.85	0.61	0.73	0.78	0.80
		0.4	0.65	0.78	0.83	0.84	0.69	0.79	0.83	0.85	0.63	0.74	0.78	0.81
		0.7	0.67	0.79	0.84	0.85	0.71	0.78	0.83	0.85	0.64	0.75	0.79	0.81
	4	0	0.75	0.83	0.89	0.89	0.73	0.85	0.89	0.91	0.69	0.83	0.87	0.89
		0.4	0.77	0.84	0.89	0.89	0.76	0.87	0.90	0.92	0.73	0.85	0.88	0.90
		0.7	0.76	0.84	0.89	0.89	0.82	0.89	0.90	0.92	0.76	0.85	0.89	0.90
	Total	0	0.95	0.97	0.98	0.98	0.94	0.97	0.98	0.98	0.94	0.97	0.98	0.98
		0.4	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.98
		0.7	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.98
100	1	0	0.74	0.87	0.89	0.92	0.77	0.85	0.87	0.89	0.80	0.89	0.92	0.93
		0.4	0.78	0.88	0.90	0.92	0.78	0.87	0.88	0.90	0.81	0.90	0.92	0.94
		0.7	0.79	0.87	0.90	0.92	0.79	0.88	0.88	0.90	0.79	0.89	0.92	0.94
	2	0	0.65	0.79	0.83	0.85	0.63	0.77	0.84	0.86	0.66	0.79	0.84	0.87
		0.4	0.67	0.80	0.83	0.86	0.65	0.78	0.85	0.87	0.68	0.79	0.84	0.88
		0.7	0.68	0.81	0.84	0.86	0.66	0.79	0.85	0.87	0.69	0.78	0.83	0.87
	3	0	0.63	0.78	0.85	0.87	0.59	0.74	0.83	0.86	0.60	0.76	0.82	0.86
		0.4	0.64	0.78	0.85	0.88	0.64	0.77	0.83	0.86	0.67	0.77	0.83	0.86
		0.7	0.64	0.77	0.85	0.88	0.67	0.77	0.83	0.86	0.67	0.78	0.82	0.86
	4	0	0.74	0.87	0.90	0.91	0.74	0.86	0.90	0.92	0.77	0.87	0.91	0.93
		0.4	0.76	0.87	0.90	0.92	0.77	0.86	0.91	0.92	0.81	0.90	0.92	0.93
		0.7	0.76	0.86	0.90	0.92	0.81	0.87	0.91	0.93	0.83	0.87	0.92	0.93
	Total	0	0.94	0.97	0.98	0.98	0.94	0.97	0.98	0.99	0.95	0.98	0.98	0.99
		0.4	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.99	0.96	0.98	0.98	0.99
		0.7	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.99	0.96	0.98	0.98	0.99

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,

ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.

Total means all ability levels.

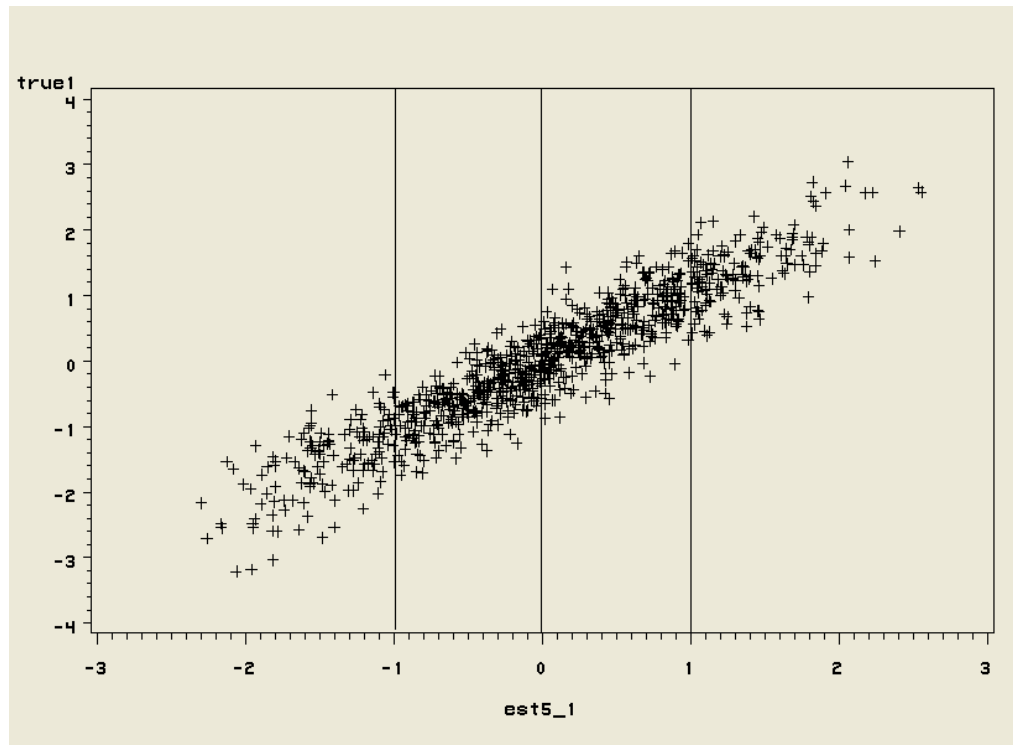


Figure 4.10. Scatter plot for all ability groups

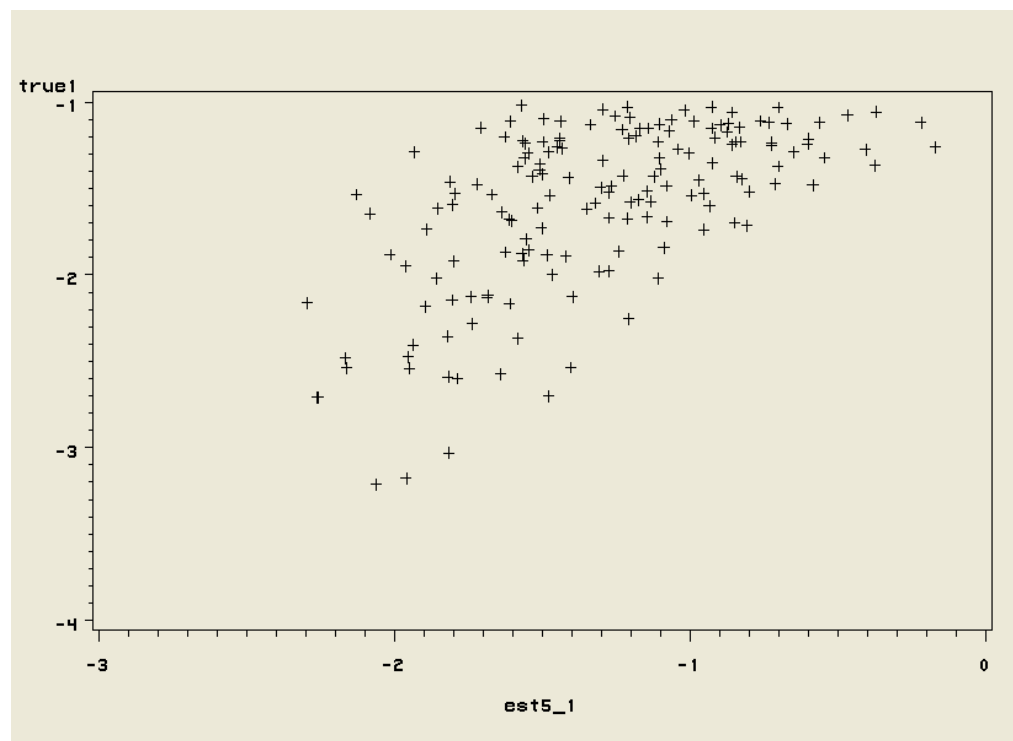


Figure 4.11. Scatter plot for examinees with true ability < -1

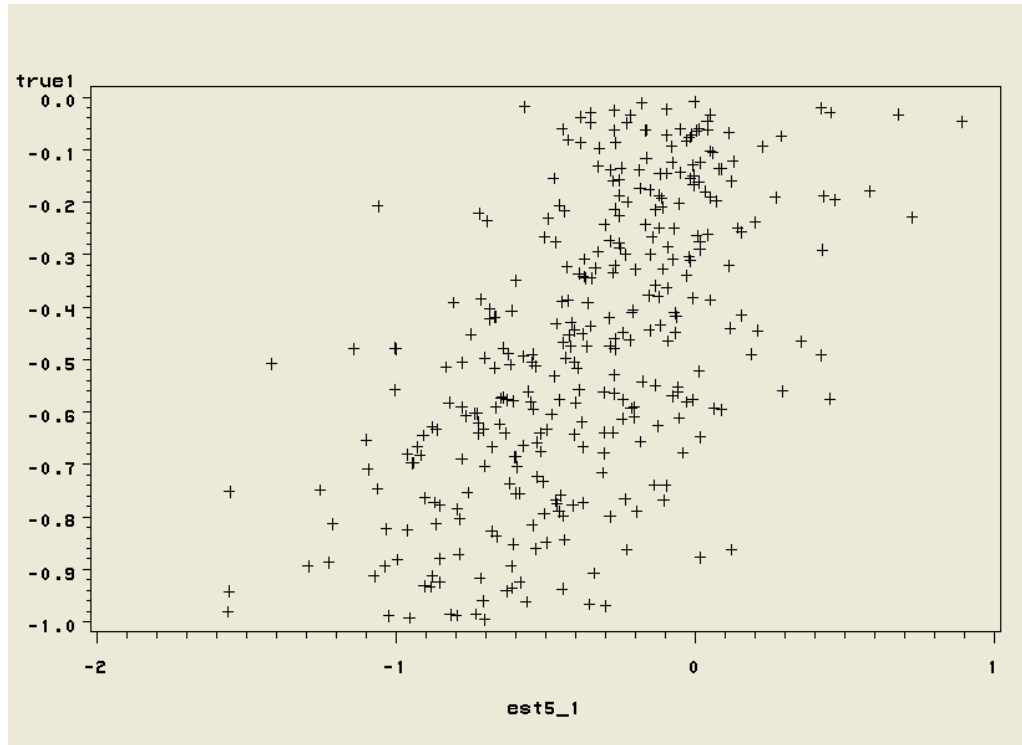


Figure 4.12. Scatter plot for examinees with $-1 < \text{true ability} < 0$

As shown in Table 4.5, the Pearson correlation between true and estimated abilities tended to increase when the correlations between dimensions increased under different item pool size, different test length, and different ability group. This was more evident when item pool size was small (10 items) and test length was short (5 items). When test length was long (20 items) and item pool size was large (100 items), the Pearson correlation tended to remain stable. This is consistent with Segall's result (1996).

Both test length and item pool size had a modest impact on the Pearson correlation. A longer test and large item pool size was associated with a small increase in the Pearson correlation results under different correlations between dimensions. As an example, the Pearson correlation on dimension 1 increased from 0.95 (test length = 5) to 0.98 (test length = 20) when item pool size = 100 and correlations = 0.4 for all examinees. As another example, the Pearson correlation on dimension 1 increased from 0.93 (item pool size = 10) to 0.94 (item pool size = 100) when test length = 5 and correlation = 0.0 for all examinees. The increase was more evident when test length was short (5 items or 10 items) and item pool size was small (10 items or 20

items). When the test length was longer (15 items or 20 items) and item pool size was large (50 items or 100 items), the Pearson correlations were relatively stable. As an example, the Pearson correlation increased from 0.94 (test length = 5) to 0.96 (test length = 10) when item pool size = 20 and correlations = 0.4 for all examinees. No gain in the Pearson correlation was observed when the test length increased from 15 items (Pearson correlation = 0.97) to 20 items (Pearson correlation = 0.97) for all examinees. Finally, the number of dimensions for ability estimates didn't provide a significant impact on Pearson correlation as shown for the RMSE and bias outcome measures.

4.1.4 Intra-class Correlation

The Pearson correlation is based on rank order. Sometimes, data may agree on ordering but not on magnitude. Therefore, an intra-class correlation (see Equation 3.4) was also calculated to check the data agreement on both rank order and magnitude.

Similar to the results for the Pearson correlations between true ability and estimated ability, the intra-class correlations between true ability and estimated ability were calculated under different item pool size, test length, correlations between dimensions, and different ability groups. The results for each combination are shown in Table 4.6.

As shown in Table 4.6, the results for the intra-class correlations were rather similar to those for the Pearson correlations. Intra-class correlations between true and estimated abilities tended to increase when the correlations between dimensions increased under different item pool sizes and different test lengths. This was more evident when item pool size was small (10 items) and test length was short (5 items). When test length was long (20 items) and item pool size was large (100 items), the intra-class correlations were also relatively stable.

As with the Pearson correlations, both test length and item pool size had an impact on intra-class correlations. A longer test and large item pool size were associated with an increase in the intra-class correlations under different correlations between dimensions. The increase was more significant when test length was short (5 items or 10 items) and item pool size was small (10 items or 20 items). When the test length was longer (15 items or 20 items) and item pool size was larger (50 items or 100 items), the intra-class correlations were stable.

Finally, as found with the RMSE and bias outcome measures, there appeared to be no impact due to the number of dimensions used to estimate ability.

Table 4.6. Intra-class Correlation

Pool Size	Ability Level	Correlation	Dimension											
			1				2				3			
			Test Length				Test Length				Test Length			
			5	10	15	20	5	10	15	20	5	10	15	20
10	1	0	0.61	0.66			0.63	0.65			0.65	0.70		
		0.4	0.67	0.70			0.67	0.69			0.70	0.74		
		0.7	0.70	0.73			0.71	0.73			0.72	0.75		
	2	0	0.55	0.59			0.54	0.58			0.55	0.60		
		0.4	0.58	0.61			0.56	0.59			0.57	0.61		
		0.7	0.59	0.62			0.57	0.61			0.58	0.63		
	3	0	0.60	0.63			0.58	0.63			0.54	0.57		
		0.4	0.61	0.64			0.60	0.65			0.56	0.58		
		0.7	0.62	0.65			0.63	0.67			0.58	0.59		
	4	0	0.68	0.70			0.61	0.65			0.69	0.73		
		0.4	0.70	0.73			0.66	0.68			0.70	0.73		
		0.7	0.72	0.75			0.67	0.69			0.69	0.72		
	Total	0	0.92	0.94			0.92	0.93			0.93	0.94		
		0.4	0.93	0.95			0.93	0.94			0.94	0.94		
		0.7	0.94	0.95			0.93	0.94			0.94	0.95		
20	1	0	0.77	0.84	0.83	0.83	0.73	0.81	0.81	0.83	0.73	0.78	0.81	0.81
		0.4	0.79	0.85	0.85	0.85	0.75	0.82	0.84	0.84	0.75	0.80	0.82	0.83
		0.7	0.81	0.85	0.85	0.86	0.75	0.82	0.84	0.84	0.75	0.81	0.83	0.83
	2	0	0.60	0.72	0.75	0.75	0.58	0.69	0.73	0.73	0.64	0.70	0.73	0.74
		0.4	0.62	0.72	0.75	0.76	0.59	0.70	0.73	0.74	0.66	0.72	0.74	0.74
		0.7	0.63	0.73	0.76	0.76	0.59	0.70	0.73	0.74	0.67	0.73	0.75	0.76
	3	0	0.59	0.69	0.73	0.73	0.66	0.74	0.76	0.78	0.57	0.67	0.71	0.72
		0.4	0.60	0.69	0.73	0.73	0.67	0.75	0.77	0.78	0.58	0.69	0.72	0.72
		0.7	0.61	0.69	0.73	0.74	0.68	0.75	0.77	0.79	0.61	0.69	0.73	0.73
	4	0	0.65	0.78	0.79	0.79	0.69	0.79	0.80	0.80	0.72	0.80	0.81	0.82
		0.4	0.68	0.78	0.80	0.81	0.72	0.79	0.80	0.81	0.75	0.81	0.82	0.84
		0.7	0.70	0.79	0.81	0.81	0.73	0.79	0.80	0.80	0.76	0.80	0.82	0.83
	Total	0	0.93	0.95	0.96	0.96	0.94	0.96	0.97	0.97	0.94	0.96	0.97	0.97
		0.4	0.93	0.96	0.97	0.97	0.95	0.96	0.97	0.97	0.95	0.96	0.97	0.97
		0.7	0.94	0.96	0.97	0.97	0.95	0.96	0.97	0.97	0.95	0.96	0.97	0.97

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,

ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.

Total means all ability levels.

Table 4.6. Continued

Pool Size	Ability Level	Correlation	Dimension											
			1				2				3			
			Test Length				Test Length				Test Length			
			5	10	15	20	5	10	15	20	5	10	15	20
50	1	0	0.75	0.83	0.86	0.88	0.65	0.78	0.83	0.87	0.77	0.87	0.90	0.91
		0.4	0.78	0.85	0.87	0.90	0.68	0.79	0.84	0.87	0.79	0.88	0.90	0.92
		0.7	0.78	0.85	0.87	0.90	0.73	0.80	0.84	0.87	0.81	0.88	0.91	0.92
	2	0	0.63	0.77	0.81	0.83	0.59	0.76	0.82	0.85	0.64	0.77	0.82	0.84
		0.4	0.64	0.76	0.81	0.83	0.60	0.77	0.82	0.85	0.65	0.77	0.82	0.84
		0.7	0.63	0.76	0.81	0.83	0.64	0.77	0.82	0.85	0.65	0.77	0.83	0.84
	3	0	0.60	0.75	0.82	0.84	0.64	0.76	0.81	0.83	0.57	0.71	0.76	0.80
		0.4	0.62	0.77	0.82	0.84	0.64	0.76	0.81	0.84	0.59	0.73	0.77	0.80
		0.7	0.65	0.78	0.83	0.85	0.67	0.76	0.82	0.84	0.60	0.74	0.78	0.80
	4	0	0.75	0.83	0.89	0.89	0.73	0.85	0.89	0.91	0.69	0.82	0.87	0.88
		0.4	0.76	0.84	0.89	0.89	0.76	0.87	0.90	0.92	0.72	0.85	0.88	0.90
		0.7	0.76	0.84	0.89	0.89	0.81	0.89	0.90	0.92	0.75	0.85	0.88	0.90
	Total	0	0.94	0.97	0.98	0.98	0.94	0.97	0.98	0.98	0.94	0.97	0.98	0.98
		0.4	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.98
		0.7	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.98
100	1	0	0.74	0.86	0.89	0.92	0.77	0.85	0.87	0.89	0.80	0.88	0.92	0.93
		0.4	0.78	0.88	0.90	0.92	0.78	0.86	0.88	0.90	0.81	0.89	0.92	0.93
		0.7	0.79	0.87	0.90	0.92	0.79	0.88	0.88	0.90	0.79	0.89	0.92	0.93
	2	0	0.62	0.78	0.83	0.84	0.59	0.75	0.83	0.86	0.62	0.77	0.82	0.86
		0.4	0.65	0.79	0.82	0.85	0.61	0.76	0.84	0.86	0.65	0.77	0.83	0.87
		0.7	0.66	0.79	0.83	0.85	0.63	0.77	0.84	0.86	0.66	0.76	0.82	0.87
	3	0	0.61	0.77	0.84	0.87	0.55	0.73	0.82	0.86	0.57	0.74	0.81	0.85
		0.4	0.62	0.77	0.85	0.87	0.61	0.75	0.83	0.86	0.63	0.75	0.81	0.85
		0.7	0.62	0.76	0.85	0.87	0.65	0.75	0.82	0.86	0.63	0.75	0.81	0.85
	4	0	0.74	0.87	0.90	0.91	0.74	0.86	0.90	0.92	0.77	0.87	0.91	0.93
		0.4	0.75	0.86	0.90	0.92	0.77	0.86	0.90	0.92	0.80	0.89	0.92	0.93
		0.7	0.76	0.85	0.90	0.92	0.81	0.87	0.91	0.93	0.83	0.87	0.92	0.93
	Total	0	0.94	0.97	0.98	0.98	0.94	0.97	0.98	0.98	0.95	0.97	0.98	0.99
		0.4	0.94	0.97	0.98	0.98	0.95	0.97	0.98	0.99	0.96	0.98	0.98	0.99
		0.7	0.95	0.97	0.98	0.98	0.95	0.97	0.98	0.99	0.95	0.98	0.98	0.99

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,
ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.
Total means all ability levels.

4.1.5 Standard Error of estimate

The standard error of estimate is defined in Equation 3.48. The amount of information (I) in Equation 3.48 was calculated as the determinant of information matrix, and was defined by Equation 3.28 and Equation 3.45. For a three-dimension CAT, the relationship between the information calculated from Equation 3.45 and the information calculated for one test was

$$I \propto I_{one\ test}^3 \quad (4.3)$$

This didn't affect the item selection process. But when the standard error of estimates were calculated, the information was calculated as

$$I = \sqrt[3]{I} \quad (4.4)$$

In order to verify this equation, 1000 examinees' responses were simulated by MCAT based on a three-dimension test with 0 correlation between dimensions and 10 items per dimension. These examinees' abilities were estimated by MCAT and MULTILOG respectively. For MCAT, the three-dimension test was counted as one test and the standard errors of estimates were calculated as above. MULTILOG was used to estimate examinees' abilities on each dimension and the standard errors of estimates were calculated for each dimension separately. The results are presented in Table 4.7. As shown in Table 4.7, the mean of standard error of estimates for the 1000 examinees which was calculated from MCAT was very similar as those calculated from MULTILOG.

Table 4.7. Standard Error of Ability Estimates

	SE across all 3 dimensions	SE from MULTILOG		
	from MCAT	Dimension 1	Dimension 2	Dimension 3
Min	0.35	0.27	0.28	0.27
Mean	0.37	0.35	0.37	0.34
Max	0.51	0.97	1.0	0.94

The standard error of estimate on all three dimensions was calculated simultaneously for each examinee. In doing so, the relationships between the dimensions were factored into the estimates of standard errors for ability estimates on all three dimensions. Therefore, the number of estimated dimensions was not considered as an independent factor for this outcome measure. Four factors were considered in these analyses: item pool size, test length, ability group, and correlations between dimensions. The values of standard error of estimates for all examinees are presented in Table 4.8. The values of standard error of estimates for the different ability groups are presented in Table 4.9 to Table 4.11 for the different dimensions. Note that each examinee could have true abilities on the different dimensions and therefore fall within different ability groups. Figures 4.13, 4.14, and 4.15 present the curves for the standard errors of ability estimates for all examinees under different correlations between dimensions, different test lengths, and different item pool sizes. The “n” on the horizontal axes is the label for each combination, and may be translated using Table 4.12.

Table 4.8. Standard Error of Estimates for all examinees

Pool Size	Correlation	Test Length			
		5	10	15	20
10	0	0.41	0.37		
	0.4	0.40	0.37		
	0.7	0.37	0.35		
20	0	0.37	0.30	0.28	0.28
	0.4	0.36	0.30	0.28	0.27
	0.7	0.34	0.29	0.27	0.26
50	0	0.36	0.27	0.23	0.21
	0.4	0.35	0.26	0.22	0.21
	0.7	0.33	0.25	0.22	0.20
100	0	0.37	0.26	0.22	0.19
	0.4	0.36	0.26	0.22	0.19
	0.7	0.34	0.25	0.21	0.19

Table 4.9. Ability level based on dimension 1

Pool Size	Ability	Correlation	Test Length			
			5	10	15	20
10	1	0.00	0.41	0.38		
		0.40	0.40	0.38		
		0.70	0.37	0.36		
	2	0.00	0.39	0.36		
		0.40	0.38	0.35		
		0.70	0.36	0.33		
	3	0.00	0.40	0.37		
		0.40	0.40	0.36		
		0.70	0.37	0.34		
	4	0.00	0.44	0.41		
		0.40	0.43	0.40		
		0.70	0.40	0.38		
20	1	0.00	0.35	0.30	0.29	0.28
		0.40	0.35	0.30	0.29	0.28
		0.70	0.33	0.29	0.27	0.27
	2	0.00	0.35	0.29	0.27	0.26
		0.40	0.35	0.29	0.27	0.26
		0.70	0.33	0.28	0.26	0.25
	3	0.00	0.37	0.30	0.28	0.27
		0.40	0.36	0.30	0.28	0.27
		0.70	0.34	0.29	0.27	0.26
	4	0.00	0.41	0.33	0.31	0.31
		0.40	0.40	0.33	0.31	0.31
		0.70	0.37	0.32	0.30	0.29
50	1	0.00	0.33	0.26	0.22	0.21
		0.40	0.33	0.26	0.22	0.21
		0.70	0.31	0.25	0.22	0.20
	2	0.00	0.33	0.25	0.21	0.20
		0.40	0.33	0.25	0.21	0.20
		0.70	0.31	0.24	0.20	0.19
	3	0.00	0.36	0.27	0.22	0.21
		0.40	0.36	0.26	0.22	0.21
		0.70	0.34	0.26	0.22	0.20
	4	0.00	0.40	0.30	0.26	0.24
		0.40	0.39	0.30	0.26	0.23
		0.70	0.37	0.28	0.25	0.23
100	1	0.00	0.34	0.25	0.21	0.19
		0.40	0.34	0.25	0.21	0.19
		0.70	0.32	0.24	0.21	0.19
	2	0.00	0.35	0.25	0.21	0.18
		0.40	0.34	0.25	0.20	0.18
		0.70	0.32	0.24	0.20	0.17
	3	0.00	0.38	0.27	0.22	0.19
		0.40	0.37	0.27	0.22	0.19
		0.70	0.35	0.26	0.22	0.19
	4	0.00	0.41	0.30	0.25	0.22
		0.40	0.40	0.29	0.24	0.22
		0.70	0.37	0.28	0.24	0.21

Table 4.10. Ability level based on dimension 2

Pool Size	Ability	Correlation	Test Length			
			5	10	15	20
10	1	0.00	0.40	0.38		
		0.40	0.40	0.38		
		0.70	0.37	0.35		
	2	0.00	0.39	0.36		
		0.40	0.38	0.35		
		0.70	0.36	0.33		
	3	0.00	0.41	0.37		
		0.40	0.40	0.36		
		0.70	0.37	0.34		
	4	0.00	0.44	0.41		
		0.40	0.44	0.40		
		0.70	0.40	0.38		
20	1	0.00	0.36	0.30	0.29	0.28
		0.40	0.35	0.30	0.28	0.28
		0.70	0.33	0.29	0.27	0.27
	2	0.00	0.35	0.29	0.27	0.26
		0.40	0.34	0.29	0.27	0.26
		0.70	0.33	0.28	0.26	0.25
	3	0.00	0.37	0.30	0.28	0.27
		0.40	0.36	0.30	0.28	0.27
		0.70	0.34	0.29	0.27	0.26
	4	0.00	0.41	0.34	0.32	0.31
		0.40	0.40	0.33	0.31	0.31
		0.70	0.38	0.32	0.30	0.29
50	1	0.00	0.33	0.26	0.22	0.21
		0.40	0.33	0.25	0.22	0.21
		0.70	0.31	0.25	0.22	0.20
	2	0.00	0.34	0.25	0.21	0.20
		0.40	0.33	0.25	0.21	0.20
		0.70	0.31	0.24	0.20	0.19
	3	0.00	0.36	0.27	0.23	0.21
		0.40	0.36	0.27	0.22	0.21
		0.70	0.34	0.26	0.22	0.20
	4	0.00	0.40	0.30	0.26	0.24
		0.40	0.40	0.30	0.26	0.24
		0.70	0.37	0.28	0.25	0.23
100	1	0.00	0.34	0.25	0.21	0.19
		0.40	0.34	0.25	0.21	0.19
		0.70	0.32	0.24	0.21	0.19
	2	0.00	0.35	0.25	0.21	0.18
		0.40	0.34	0.25	0.20	0.18
		0.70	0.32	0.24	0.20	0.17
	3	0.00	0.38	0.27	0.22	0.19
		0.40	0.37	0.27	0.22	0.19
		0.70	0.35	0.26	0.22	0.19
	4	0.00	0.42	0.30	0.25	0.22
		0.40	0.40	0.29	0.25	0.22
		0.70	0.37	0.28	0.24	0.21

Table 4.11. Ability level based on dimension 3

Pool Size	Ability	Correlation	Test Length			
			5	10	15	20
10	1	0.00	0.40	0.38		
		0.40	0.40	0.38		
		0.70	0.37	0.35		
	2	0.00	0.39	0.36		
		0.40	0.38	0.35		
		0.70	0.36	0.33		
	3	0.00	0.41	0.37		
		0.40	0.40	0.36		
		0.70	0.37	0.34		
	4	0.00	0.44	0.41		
		0.40	0.43	0.40		
		0.70	0.40	0.38		
20	1	0.00	0.35	0.30	0.29	0.28
		0.40	0.35	0.30	0.29	0.28
		0.70	0.33	0.29	0.27	0.27
	2	0.00	0.35	0.29	0.27	0.26
		0.40	0.35	0.29	0.27	0.26
		0.70	0.33	0.28	0.26	0.25
	3	0.00	0.37	0.30	0.28	0.27
		0.40	0.36	0.30	0.27	0.27
		0.70	0.34	0.28	0.26	0.26
	4	0.00	0.41	0.34	0.32	0.31
		0.40	0.40	0.33	0.31	0.31
		0.70	0.38	0.32	0.30	0.29
50	1	0.00	0.33	0.26	0.22	0.21
		0.40	0.33	0.25	0.22	0.21
		0.70	0.31	0.25	0.22	0.20
	2	0.00	0.34	0.25	0.21	0.20
		0.40	0.33	0.25	0.21	0.20
		0.70	0.31	0.24	0.20	0.19
	3	0.00	0.36	0.27	0.23	0.21
		0.40	0.36	0.27	0.22	0.21
		0.70	0.34	0.26	0.22	0.20
	4	0.00	0.40	0.30	0.26	0.24
		0.40	0.40	0.30	0.26	0.24
		0.70	0.37	0.29	0.25	0.23
100	1	0.00	0.34	0.25	0.21	0.19
		0.40	0.33	0.25	0.21	0.19
		0.70	0.31	0.24	0.21	0.19
	2	0.00	0.35	0.25	0.21	0.18
		0.40	0.34	0.25	0.20	0.18
		0.70	0.32	0.24	0.20	0.17
	3	0.00	0.38	0.27	0.22	0.19
		0.40	0.37	0.27	0.22	0.19
		0.70	0.35	0.26	0.22	0.19
	4	0.00	0.41	0.30	0.25	0.22
		0.40	0.40	0.29	0.24	0.21
		0.70	0.37	0.28	0.24	0.21

Table 4.12. Interpretation of N in Figure 4.13 – Figure 4.15

N	Figure 4.13		Figure 4.14		Figure 4.15	
	pool	length	pool	correlation	correlation	length
1	10	5	10	0	0	5
2	10	10	10	0.4	0	10
3	10	15	10	0.7	0	15
4	10	20	20	0	0	20
5	20	5	20	0.4	0.4	5
6	20	10	20	0.7	0.4	10
7	20	15	50	0	0.4	15
8	20	20	50	0.4	0.4	20
9	50	5	50	0.7	0.7	5
10	50	10	100	0	0.7	10
11	50	15	100	0.4	0.7	15
12	50	20	100	0.7	0.7	20
13	100	5				
14	100	10				
15	100	15				
16	100	20				

As shown in Table 4.8 and Figure 4.13, the impact of correlations between dimensions on the standard errors was observable. The standard error of estimates tended to decrease when correlations increased for all cases, especially when item pool size was small (10 items), as shown in Figure 4.13 where $n < 5$. When item pool size was large (50 items or 100 items) and test length was long (15 items or 20 items), standard error of estimates showed similar values no matter what the size of the correlation.

Larger item pool sizes and longer test lengths were also associated with decreases in the standard errors of ability estimates. As indicated in Figure 4.14, line 1 (test length = 5) had the largest standard error of estimates, followed by line 2 (test length = 10), line 3 (test length = 15), and line 4 (test length = 20). Similarly in Figure 4.15, line 1 (item pool size = 10) had the largest standard error of estimates, followed by line 2 (item pool size = 20), line 3 (item pool size = 50), and line 4 (test item pool size = 100). The difference in standard error of estimates was more pronounced when item pool size was small (10 items or 20 items, see Figure 4.14) or test length was short (5 items or 10 items, see Figure 4.15). When item pool size was large (50 items or 100 items) or test length was long (15 items or 20 items), the difference was negligible.

A mixed ANOVA was performed on the standard errors of ability estimates as a function of item pool size (10, 20, 50, 100), correlations between dimensions (0.0, 0.4, 0.7), and test length (5, 10, 15, 20). Item pool size and correlations between dimensions were between-subject factors and test length was within-subject factor. Note that for these analyses, the ability group factor was again not incorporated into the analysis since simulated examinees were not assigned to one ability group for all three dimensions. As such, the ability group factor was not “crossed” with the other factors.

Table 4.13 shows the results of the mixed ANOVA. Using $\alpha = .05$, all effects were statistically significant, including three main effects, three two-way interaction effects, and one three-way interaction effect. Using η^2 as the measure of effect size, the main effect of test length accounted for most of the variance (93%). As before, a number of significant interaction terms exhibited zero effect sizes. The one interaction that may be of practical significance is the length*item pool interaction. This interaction can be observed in Figure 4.15 where a longer test length effect was observed in larger item pools.

Table 4.13. ANOVA Results

Source	df	F	P	η^2
Length	1.279	527949.5	0	.93
Length * Pool	2.557	16180.54	0	.05
Length * Correlation	2.557	1169.40	0	0
Length * Pool * Correlation	5.120	21.899	0	0
Pool	2	3696.459	0	0
Correlation	2	331.858	0	0
Pool * Correlation	4	3.785	.004	0
$R^2 = 0.99$				

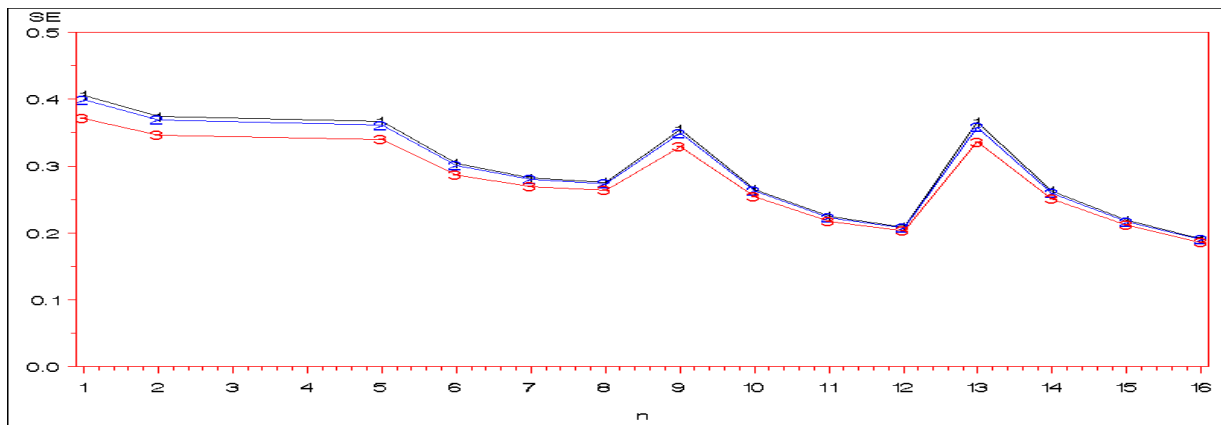


Figure 4.13. SE of estimates under different correlations between dimensions
(line1: correlation =0.0, line2: correlation =0.4, line3: correlation =0.7)

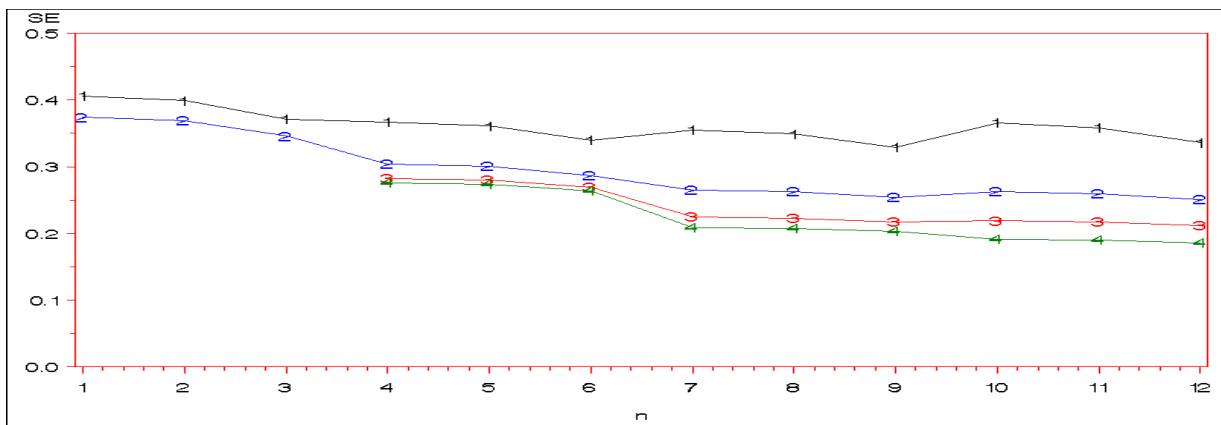


Figure 4.14. SE of estimates under different test length
(line1: length =5, line2: length =10, line3: length =15, line4: length =20)

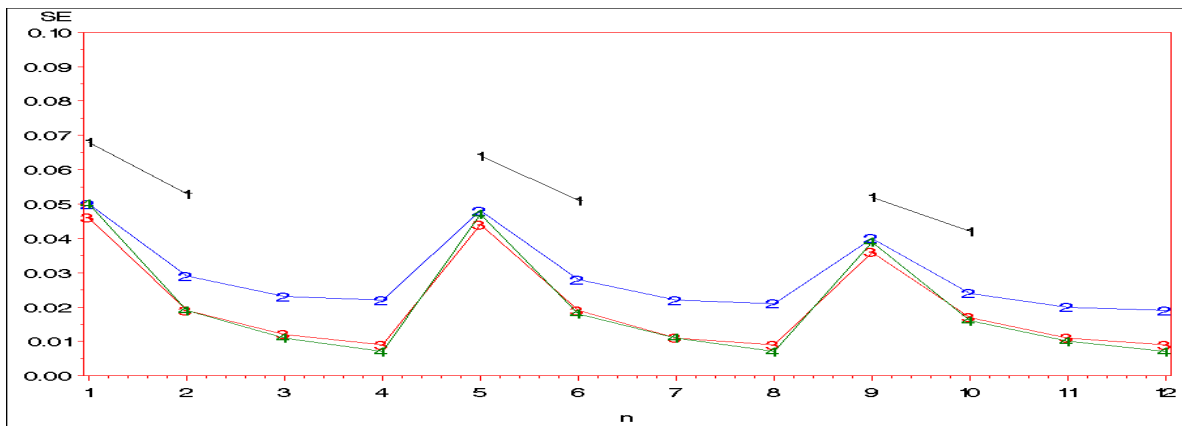


Figure 4.15. SE of estimates under different item pool sizes
(line1: pool =10, line2: pool =20, line3: pool =50, line4: pool = 100)

4.1.6 Optimal size of the item pool

Based on the previous sections, an increase in the item pool size provides improvement in most outcome measures, including RMSE, Pearson correlations, intra-class correlations, and standard error of estimates. However, the item writing costs are linear with item pool size. When the item pool size is large enough, the improvement could be trivial and not cost effective.

The size of the item pool depends on the intended purpose and characteristics of the tests being constructed. Weiss (1995) points out that satisfactory implementation of CAT has been obtained with an item pool of 100 high quality, well distributed items. He also notes that properly constructed item pools with 150-200 items are preferred. If one is going to incorporate a realistic set of constraints (e.g. random selection from among the most informative items to minimize item exposure; selection from within sub-skills to provide content balance) or administer a very high stake examination, then a much larger item pool would be needed. This study was a simple case, because item exposure is not an issue for surveys and test length for each dimension was the same.

Current CAT test administration methods fall into two basic categories. These two types of CAT are defined by their stopping rules: fixed-length and variable-length tests. A fixed-length CAT administers the same number of items to each examinee. Different examinees therefore may be tested to different levels of precision. In contrast, a variable-length CAT tests each examinee to a fixed level of precision, relative to a desired standard error of ability estimates, even if this requires administering different numbers of items to different examinees. The optimal item pool size therefore may be different based on the stopping rule. For fixed-length tests, all examinees have same number of items, but their standard errors of estimates are different. And the standard error of estimates tends to decrease when item pool size increases, as shown in Figure 4.16. When item pool size is big enough, the curve converges. For variable-length tests, when the standard error of estimated is smaller than the desired standard errors for ability estimates, the tests stop. Therefore, examinees could have different test lengths, and the test length tends to decrease when item pool size increases, as shown in Figure 4.17. Similar to fixed-length tests, when the item pool size is large enough, the curve converges. The elbow at which the curves converge reflects the optimal item pool size.

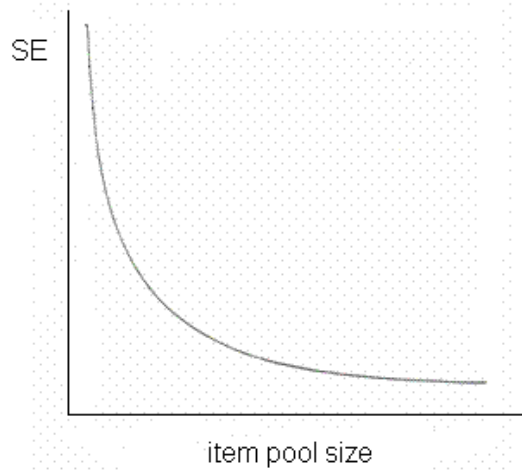


Figure 4.16. Optimal Pool Size Curves for fixed-length tests

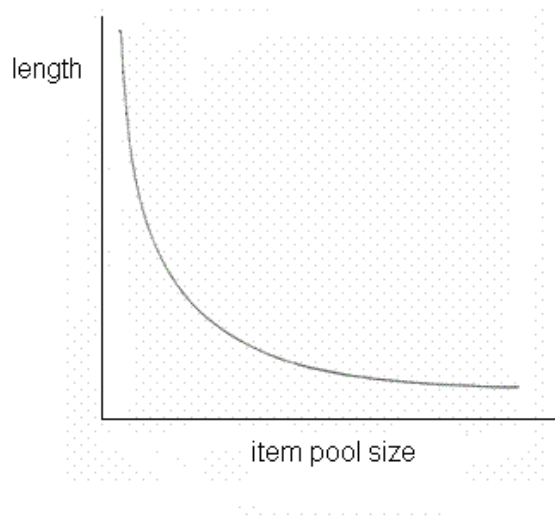


Figure 4.17. Optimal Pool Size Curves for variable-length tests

This study is based on fixed-length tests. The optimal item pool size curves are shown in Figure 4.18 through Figure 4.20. Only the standard error of the test, RMSE, and Pearson correlation are presented here because positive values and negative values cancel each other for bias and intra-class correlations have rather similar results from the Pearson correlation. The curves for the Pearson correlation and RMSE are presented only for dimension 1 because the

curves for dimension 2 and 3 are very similar to that of dimension 1. The optimal item pool size is highly related to test length. Therefore, the curves were plotted under each test length.

As shown in Figure 4.18 through Figure 4.20, when test length was short (test length = 5 or 10), two elbows were observed at item pool size = 20 and 50 respectively. When test length was longer (length = 15 or 20), only one elbow was observed at item pool size = 50. After the item pool size = 50, the curves remained stable. Therefore, the optimal item pool size was 50 items for this study.

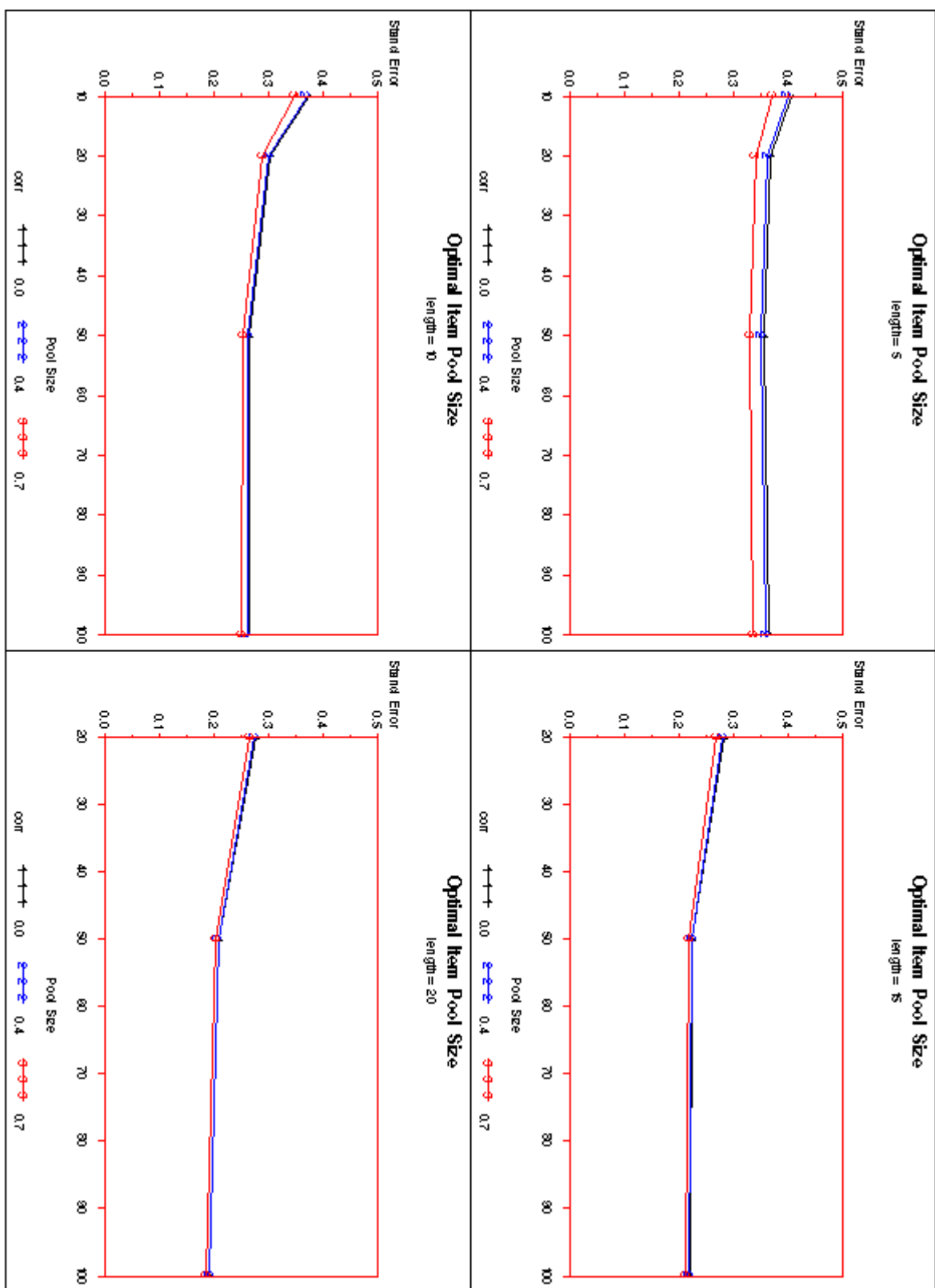


Figure 4.18. Optimal Pool Size for Standard Error of Estimates

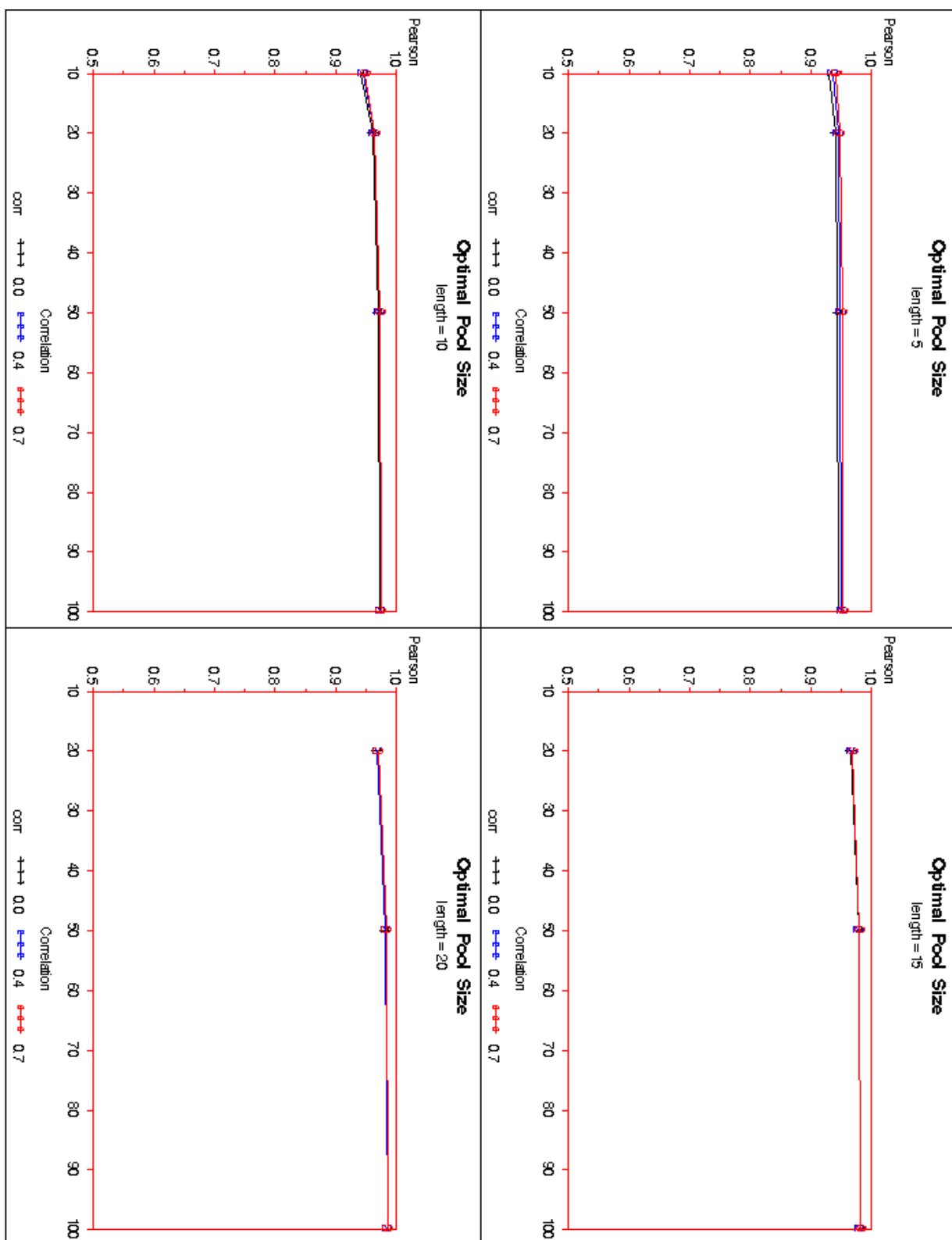


Figure 4.19. Optimal Pool Size for Pearson Correlation

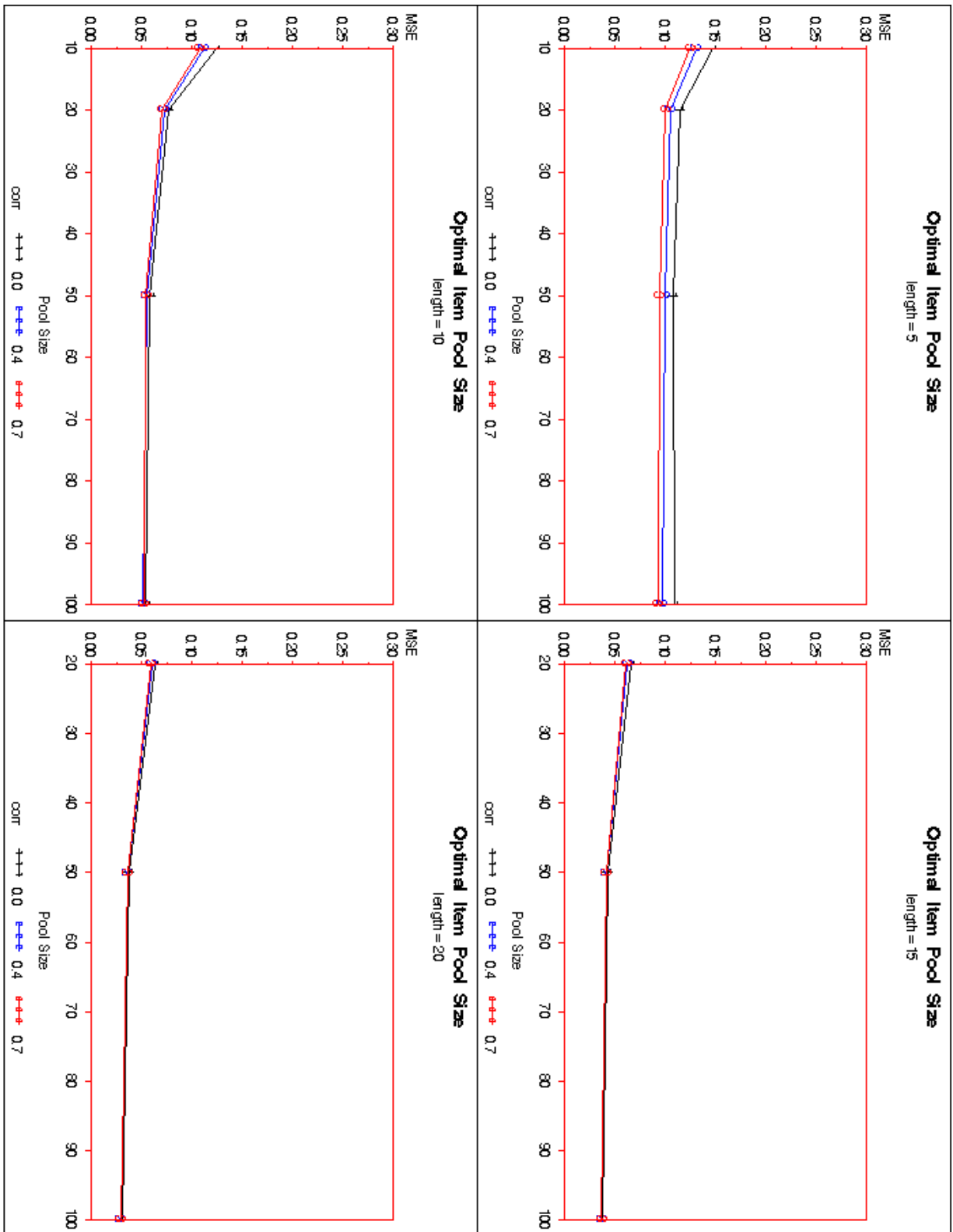


Figure 4.20. Optimal Pool Size for RMSE

4.2 RESULTS FROM REAL DATA

The application component to this study utilized real data from a uni-dimensional survey “DASH” and a two-dimensional survey “SF-36”. The SF-36 is a general measure of health status that measures eight domains of health including physical functioning, role limitation due to physical problems, bodily pain, general health, vitality, social functioning, role limitations due to emotional problems, and mental health. The SF-36 is a 36-item survey, containing 2 to 10 items for each domain. All items use Likert-type scales, the number of order categories for each item range from 2 to 6. The DASH is a 30-item self-reported questionnaire used routinely in clinical practice to assess rehabilitation outcomes for individuals with a variety of upper extremity impairments. The DASH is designed to measure physical function and symptoms. It includes 21 physical function items, six symptom items, and three social/role function items. All items use a five-category Likert-type scale.

4.2.1 Data Analyses

Before the MCAT could be simulated, item parameters and ability values required calibration.

SF-36 items and scales were scored so that a higher score indicated a better health state. In contrast, DASH items and scales were scored so that a lower score indicated a better health state. Therefore, DASH items were reverse coded to ensure that a higher item value indicated better health. 10 items from SF-36 items also required recoding.

The SF-36 was constructed to represent two major dimensions of health: physical and mental. Items for the dimensions were selected based on information from the technical manual (Ware, Kosinski, & Keller, 1994). The factor loadings are summarized in Table 4.14, and indicate that physical functioning, role limitation due to physical problems, and bodily pain had a strong correlation with factor 1 and weak correlation with factor 2; mental health, role limitations due to emotional problems, and social functioning had a strong correlation with factor 2 and weak correlation with factor 1; general health and vitality appeared to measure both. By examining the item content, factor 1 was labeled physical component and factor 2 was labeled mental component.

Table 4.14. Factor loadings for SF-36

	Physical	Mental
Physical Functioning	.88	.04
Role-Physical	.78	.30
Bodily Pain	.77	.24
Mental Health	.12	.90
Role-Emotional	.19	.81
Social Functioning	.44	.71
Vitality	.59	.57
General Health Perceptions	.68	.32

A subset of 26 items was used from the SF-36: 16 items reflect the physical function and 10 items reflect the mental function. 10 items were excluded from this study since the focus was an assessment with single structure, and these 10 items loaded on both dimensions. The item parameters were calibrated by MULTILOG (Thissen, 1991) separately for DASH and physical and mental dimensions of SF-36, and were treated as the population item parameters. Because the item parameter calibration is sensitive to the test length and extreme values can be achieved by short test length, a prior distribution was imposed when item parameters were calibrated for physical dimension and mental dimension of SF-36. The item parameters were then used to construct a three-dimension item pool for the MCAT application (see Table 4.15).

The examinees' ability values on three dimensions were estimated by MULTILOG and were treated as the true abilities. These abilities values were correlated and the correlations (as shown as below) were used as the correlations between dimensions in this study. The MCAT was conducted based on the calibrated item parameters (see Table 4.16) and ability values.

Table 4.15. Item Pool for Real Data

A1	A2	A3	B1	B2	B3	B4	B5	Dimension	# of categories
1.71	0	0	-1.34	-0.52	0.33	1.16	.	1	5
1.32	0	0	-3.38	-2.41	-1.27	-0.34	.	1	5
1.62	0	0	-2.89	-2.09	-1.27	-0.48	.	1	5
2.54	0	0	-2.17	-1.67	-0.74	0.08	.	1	5
2.25	0	0	-1.97	-1.12	-0.26	0.66	.	1	5
1.89	0	0	-1.58	-0.64	0.20	1.05	.	1	5
3.31	0	0	-0.87	-0.27	0.46	1.1	.	1	5
3.03	0	0	-0.92	-0.35	0.39	1.11	.	1	5
2.93	0	0	-1.89	-1.25	-0.37	0.35	.	1	5
2.36	0	0	-1.74	-1.13	-0.24	0.66	.	1	5
2.42	0	0	-1.16	-0.46	0.336	1.13	.	1	5
2.26	0	0	-1.14	-0.57	0.09	0.81	.	1	5
2.37	0	0	-1.9	-1.12	-0.25	0.50	.	1	5
2.17	0	0	-1.22	-0.57	0.21	1.07	.	1	5
2.03	0	0	-2.12	-1.06	-0.10	0.78	.	1	5
2.15	0	0	-2.04	-1.45	-0.69	-0.08	.	1	5
1.99	0	0	-2.2	-1.74	-1.04	-0.32	.	1	5
2.52	0	0	-0.58	0.10	0.88	1.72	.	1	5
2.57	0	0	-0.64	0.02	0.74	1.42	.	1	5
2.24	0	0	-2.07	-1.65	-0.90	-0.21	.	1	5
1.81	0	0	-1.83	-1.55	-0.95	-0.27	.	1	5
1.59	0	0	-2.34	-1.08	-0.13	0.82	.	1	5
1.94	0	0	-1.87	-0.67	0.35	1.4	.	1	5
1.43	0	0	-2.19	-0.48	1.15	2.75	.	1	5
1.62	0	0	-1.61	0.03	1.46	2.69	.	1	5
0.767	0	0	-4.36	-2.46	-0.70	0.48	.	1	5
1.54	0	0	-1.68	-0.48	0.81	1.81	.	1	5
1.3	0	0	-2.01	-0.65	0.64	1.58	.	1	5
1.24	0	0	-2.78	-1.27	0.14	1.37	.	1	5
1.16	0	0	-1.84	0.15	1.05	1.95	.	1	5
0	1.64	0	0.27	1.73	.	.	.	2	3
0	1.73	0	-0.55	0.97	.	.	.	2	3
0	1.64	0	-1.0	0.68	.	.	.	2	3
0	1.89	0	-2.05	-0.78	.	.	.	2	3
0	2.12	0	-2.45	-1.42	.	.	.	2	3
0	1.57	0	-2.09	-0.86	.	.	.	2	3
0	1.97	0	-1.67	-0.87	.	.	.	2	3
0	2.29	0	-1.90	-1.15	.	.	.	2	3
0	2.25	0	-2.40	-1.56	.	.	.	2	3
0	1.25	0	-2.46	-0.34	.	.	.	2	3
0	1.93	0	-0.01	2	2
0	2.20	0	0.32	2	2
0	2.12	0	0.55	2	2
0	2.04	0	0.62	2	2
0	1.17	0	-3.42	-1.33	0.74	1.91	3.40	2	6
0	1.54	0	-2.16	-0.76	0.45	1.79	3.17	2	6
0	0	1.88	-1.41	3	2
0	0	1.88	-1.18	3	2
0	0	1.83	-1.59	3	2
0	0	1.57	-3.05	-1.72	-0.91	0.02	.	3	5
0	0	1.40	-3.93	-2.88	-2.16	-1.06	0.08	3	6
0	0	1.91	-4.01	-3.16	-2.50	-1.66	-0.80	3	6
0	0	1.65	-2.45	-1.43	-0.36	0.43	2.26	3	6
0	0	1.85	-3.98	-2.93	-2.30	-1.21	-0.18	3	6
0	0	1.67	-3.12	-2.18	-1.03	-0.25	1.77	3	6
0	0	1.67	-2.84	-1.77	-0.75	0.03	.	3	5

4.2.2 Outcome measures

As in study 1, RMSE, bias, Pearson correlations, and intra-class correlations were calculated for each combination of ability level, test length, and correlations between dimensions. The standard errors for ability estimates were also calculated for each combination of test length and correlations between dimensions. The results are summarized in Table 4.16 through Table 4.20.

When the value of correlation between dimensions = 0.0 was compared with the value of correlations between dimensions = Φ , in general, the RMSE, the absolute value of bias, and standard error of estimates decreased when correlation between dimensions increased, while the Pearson correlations and intra-class correlations between true ability and estimated ability increased when correlations between dimensions increased. However, there were some exceptional cases. As an example, in Table 4.17, when test length = 5 and ability level = 2, on dimension 2, bias increased from 0.12 (correlation = 0.0) to 0.16 (correlation = Φ).

Consistent with the study based on simulated data and as would be expected, a longer test improved the accuracy of ability estimates. The RMSE, the absolute value of bias, and the standard error of estimates tended to decrease when test length increased under different ability level and correlation between dimensions. Pearson correlations and intra-class correlations between true ability and estimated ability tended to increase when test length increased under different correlations between dimensions. As an example, in Table 4.20, standard errors for ability estimates decreased from 0.47 (test length = 5, correlations between dimensions = 0.0) to 0.38 (test length = 10, correlation between dimensions = 0.0). Another example, in Table 4.16, on dimension 1, RMSE decreased from 0.18 (ability level = 1, correlation = 0.0, test length = 5) to 0.10 (ability level = 1, correlation = 0.0, test length = 10).

The impact of ability level on RMSE and bias was also consistent with the study based on simulated data. When ability was in the extreme range ($\theta < -1$ or $\theta > 1$), the RMSE and the absolute value of bias had larger values than those in the medium range ($-1 < \theta < 1$). As an example, in Table 4.16, when the correlation = 0.0, test length = 5, on dimension 1, the RMSE was 0.18 (ability level = 1) and 0.13 (ability level = 4), while the RMSE was 0.09 (ability level = 2) and 0.08 (ability level = 3). As indicated in Table 4.17, most bias values were positive for

ability = 1 or 2, and most bias values were negative for ability = 3 or 4, which was consistent with the study based on simulated data. This indicated that the ability was over-estimated for examinees who had negative true ability value and the ability was under-estimated for examinees who had positive true ability value.

The impact of number of dimensions for ability estimates was not notable as found in the study based on simulated data. The reason may be due to the small item pool size. There were only 16 items and 10 items on dimension 2 and dimension 3 respectively. There were not enough “good” items to be selected. Therefore, the errors in estimating ability were larger.

Table 4.16. Root Mean Square Error

Ability Level	Correlation	Test Length					
		Dimension 1		Dimension 2		Dimension 3	
		5	10	5	10	5	10
1	0	0.18	0.10	0.53	0.37	0.13	0.03
	Φ	0.16	0.09	0.57	0.34	0.10	0.03
2	0	0.09	0.05	0.23	0.20	0.14	0.03
	Φ	0.08	0.04	0.24	0.22	0.14	0.03
3	0	0.08	0.04	0.24	0.22	0.14	0.03
	Φ	0.07	0.03	0.09	0.05	0.09	0.02
4	0	0.13	0.08	0.33	0.17	0.31	0.24
	Φ	0.12	0.08	0.25	0.15	0.27	0.21

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,
ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.

Table 4.17. Bias

Ability Level	Correlation	Test Length					
		Dimension 1		Dimension 2		Dimension 3	
		5	10	5	10	5	10
1	0	0.23	0.17	0.10	-0.08	0.07	-0.08
	Φ	0.19	0.15	0.08	-0.04	0.06	-0.08
2	0	0.04	0.05	0.12	0.12	-0.07	-0.03
	Φ	0.01	0.03	0.16	0.13	-0.01	0.01
3	0	-0.05	-0.05	-0.08	0.0	-0.15	-0.10
	Φ	-0.05	-0.05	-0.02	0.05	-0.10	-0.05
4	0	-0.16	-0.16	-0.43	-0.33	-0.48	-0.46
	Φ	-0.16	-0.14	-0.37	-0.29	-0.43	-0.40

Note: ability 1 means: $\theta \leq -1$, ability 2 means: $-1 < \theta < 0$,
ability 3 means: $0 \leq \theta < 1$, ability 4 means: $\theta \geq 1$.

Table 4.18. Pearson Correlation

Correlation	Test Length					
	Dimension 1		Dimension 2		Dimension 3	
	5	10	5	10	5	10
0	0.94	0.97	0.89	0.92	0.94	0.98
Φ	0.95	0.98	0.89	0.92	0.93	0.98

Table 4.19. Intra-Class Correlation

Correlation	Test Length					
	Dimension 1		Dimension 2		Dimension 3	
	5	10	5	10	5	10
0	0.94	0.97	0.88	0.92	0.93	0.98
Φ	0.95	0.97	0.89	0.92	0.93	0.97

Table 4.20. Standard Error of Estimates

Correlation	Test Length	
	5	10
0	0.47	0.38
Φ	0.44	0.37

5.0 DISCUSSION

5.1 SUMMARY OF FINDINGS

This study aimed to compare the efficiencies of multi-dimensional CAT versus uni-dimensional CAT based on the multi-dimensional graded response model and provide information about the optimal size of the item pool. To achieve these goals, item selection and ability estimation methods based on multi-dimensional graded response models were developed and two studies, one based on simulated data, the other based on real data, were conducted. For both studies, a SAS program was developed to simulate computer adaptive testing in the context of psychological and health assessments. Five design factors were manipulated: 1) correlation between dimensions, 2) item pool size, 3) test length, 4) ability group, and 5) number of estimated dimensions. Five outcome measures, the Pearson and intra-class correlations between estimated ability and true ability, root mean squared error, bias, and standard error for trait estimates, were calculated based on three correlations between dimensions (0.0, 0.4, and 0.7), four item pool sizes (10, 20, 50, 100), four test lengths (5, 10, 15, and 20 items), and four levels of each ability ($\theta \leq -1$, $-1 < \theta < 0$, $0 \leq \theta < 1$, and $\theta \geq 1$).

A multi-dimensional CAT is more complicated when compared with a uni-dimensional CAT. The present study involved several design constraints: 1) a simple factor structure, where each item loads on a single dimension was assumed; 2) only Bayesian based methods were used since these methods considered the correlations between different dimensions; 3) only fixed-length CATs were considered since these are currently used in most CAT assessments; 4) other issues, such as exposure control, content balancing, are not typical problems for psychological and health assessment, and therefore were not considered.

The specific research questions were under study:

1. How does the correlation between dimensions affect the efficiency of multi-dimensional CAT?
2. Is a multi-dimensional CAT more efficient than a uni-dimensional CAT?
3. Is there any difference between the results at different levels of the trait?
4. What is the optimal size of the item pool?

The results of this simulation study provide evidence to answer each research question. The impact of correlation between dimensions on efficiency of multi-dimensional CAT was observed from a comparison of the outcome measures under different correlations between dimensions. A modest effect due to the correlation between dimensions on the outcome measures was observed, although the effect was found primarily for correlations of 0 versus 0.4. When the correlations between dimensions increased, the root mean squared error and the standard error of estimates tended to decrease for all three dimensions, the Pearson correlations and intra-class correlations between true and estimated abilities tended to increase.

When each item loads on a single dimension and the dimensions are uncorrelated (correlation between dimensions = 0.0), the item selection and ability estimation procedures that are based on a multi-dimensional model are equivalent to those methods based on a uni-dimensional model. Based on a comparison of this condition with conditions in which the correlation between dimensions was greater than 0, the multi-dimensional CAT was more efficient than the uni-dimensional CAT. The gains in efficiency obtained by the multi-dimensional CAT depend on the correlations between dimensions. In general, the larger the magnitude of these correlations, the higher the gains in efficiency over uni-dimensional CAT. As Segall (1996) pointed out the gain in efficiency can be attributed to: 1) item selection and 2) ability estimation. As defined in Chapter 2.3, the Bayesian-based item selection and ability estimation methods take the correlation between dimensions into account, which leads to noticeable improvements in ability estimates.

The third research question was addressed by comparing the root mean squared error and bias under different ability levels. Not unexpectedly, ability level had an impact on the outcome measures. The results indicated that, for true ability value in the middle range ($-1 < \theta < 1$), both root mean squared error and absolute value of bias were smaller than those values when true ability value was in the extreme range ($\theta < -1$, or $\theta > 1$). As would be the case for a uni-

dimensional CAT, a multi-dimensional CAT provided more accurate estimates for those examinees with average true ability values than those with true ability values in the extreme range. This was explained by examining the test information function which illustrated that more information for estimating ability was available in the middle range ($-1 < \theta < 1$) than in the extreme range ($\theta < -1$ and $\theta > 1$) and slightly more information was available at the higher ability range than the lower ability range. In addition, the direction of any bias was as expected with Bayesian estimates. When true ability was negative ($\theta < 0$), bias was positive or $\hat{\theta} > \theta$. When true ability level was positive ($\theta > 0$), all bias values were negative or $\hat{\theta} < \theta$. Therefore, for examinees with negative true ability values, their estimated ability values were over-estimated. For examinees with positive true ability values, their estimated ability values were under-estimated. This is consistent with Bayesian estimation methods which “shrink” estimates toward the mean of the prior distribution.

Information on the optimal item pool size was provided by plotting the outcome measures versus the item pool size. The plots indicated that, for short test (5 items), the optimal item pool size was 20 items; for longer test (> 5 items), the optimal item pool size was 50 items. However, if item exposure control or content balancing were an issue, a larger item pool would be needed to achieve the same efficiency in ability estimates.

The results of this simulation study provide compelling evidence for several findings as well. The first significant finding was observed from a comparison of the outcome measures for ability estimates for dimension 1, dimension 2, and dimension 3. Recall that one feature of the study was that items were administered one dimension at a time. Thus, the ability estimates for dimensions after dimension 1 were based not only on items in the administered dimension but also on the relationship between dimensions. The results showed that as the number of dimensions used for estimating ability increased, RMSE and the absolute values of bias decreased. Ability estimates for dimension 3 had the most accurate estimates, followed by dimension 2, and followed by dimension 1. This supports the idea that the ability of one dimension can be used to augment the information available to estimate ability in another dimension.

The effect of item pool size and test length on ability estimates were similar to uni-dimensional CAT (Van der Linden, 1997; Wang & Vispoel, 1998; Warm, 1989). Larger item pools and longer tests yielded more accurate and reliable estimations. The results indicated that,

as more items were in the item pool, the smaller the root mean squared error and standard errors of ability estimates, and the higher Pearson and intra-class correlations. However, the absolute values of bias measure tended to decrease when item pool size increased, although this trend was not consistent across all experimental conditions. For conditions with the same item pool size, as the test length was longer, smaller root mean squared error, absolute value of bias, and standard error of ability estimates were observed, whereas higher Pearson and intra-class correlations were observed.

In order to investigate the significance of the manipulated factors, an ANOVA was conducted for RMSE and standard errors of ability estimates. Using η^2 as the measure of effect size, the main effect of test length accounted for most of the variance for RMSE (6%) and standard errors of estimates (93%).

The application component of the study investigated the same factors as the simulation study using real data from a uni-dimensional survey “DASH” and a two-dimensional survey “SF-36”. Results of the real data application found similar effects as those observed in the simulation study: the accuracy of ability estimates was improved by a longer test or when correlations between dimensions increased; the ability estimates were more accurate for examinees with ability in the medium range ($-1 < \theta < 1$) than examinees with ability in the extreme range ($\theta < -1$ or $\theta > 1$); and the ability was over-estimated for examinees who had negative true ability value and the ability was under-estimated for examinees who had positive true ability value.

5.2 LIMITATIONS AND FUTURE RESEARCH

Future research might take a variety of directions. First, this study assumed a simple structure, where each item loads on a single dimension. This made it easy to compare a uni-dimensional CAT with a multi-dimensional CAT. However, in practice, items may provide information on more than one dimension. Future research could be conducted to assess how the borrowing of information from the other dimension affects the gain in efficiency based on a complex structure.

A second direction may be to study MCAT with more than three dimensions. The results indicated that MCAT had more accurate estimates as the number of dimensions for estimates increased. This study was limited to three dimensions. However, the gain in efficiency related to higher dimensionality could also be conducted.

It may be also of interest to examine the effects of smaller correlations between the dimensions (e.g., .2 - .3) as well as larger correlations between dimensions (e.g., .8 - .9). These correlations would consider assessment that are less or more strongly correlated respectively.

This study also assumed fixed-length tests. Examinees were administered the same number items from each dimension and the subtests were administered according to a fixed order. For example, if test length was 10 items, 10 items from dimension 1 were selected first, followed by 10 items from dimension 2, and then followed by 10 items from dimension 3. Whether administering subtests according to a fixed order for each examinee or administering items to examinees from whatever subtest affords the most information may be an important factor that needs to be investigated in future research. This issue doesn't refer to cases in which no items from dimension would be administered due to lack of efficient information. Rather, the question is about selecting items from any dimensions would improve the efficiency of the CAT.

This study focused on comparing the multi-dimensional and uni-dimensional computer adaptive assessment. Although issues, such as item exposure control methods, item context effects, or content balancing, may not be relevant to all psychological and health assessment, they may be relevant to some applications. All these issues have been well documented for uni-dimensional CAT. However, the research based on multi-dimensional CAT is not known. Therefore, further study is needed to investigate these problems in the multi-dimensional context. The optimal item pool size could also be explored when considering these issues.

Another potential direction for future research is related to the order in which dimension are administered. It is possible that different dimension afford different amounts of information about ability. Thus, it may be more efficient to administer the dimension that affords the most information followed by the dimension that affords less information. The order of dimensions could be directly manipulated in a future study.

Finally, more efficient numerical methods could be developed to improve the running time of computer software. For this study, a SAS program, MCAT, was developed to simulate a computer adaptive test based on the multi-dimensional graded response model. The ability

values on three dimensions were estimated simultaneously based on the Newton-Raphson procedure. The running time was much longer when three-dimension ability values were estimated than that of the uni-dimensional test. With more dimensions, running time would be even longer. Therefore, more efficient numerical methods may be needed before multi-dimensional computer adaptive tests can be routinely applied.

APPENDIX A

SAS PROGRAM TO SIMULATE MCAT

```

*****;
* CAT_SIMULATOR
* SAS program for administering a CAT;
* CAT can use simulated item responses or select responses from existing data;
*****;

options mprint mlogic;

* USER CONTROL VARIABLES;

%let nExaminee=1000;

%let parmfile='parameter\par10.par'; * parameter file which is real value;

%let outputfile = '\output\out_0.0.dat'; /*record estimated ability,admin items,provSE*/

%let responsefile = '\output\response_0.0.dat';

%let adminfile = '\output\admin_0.0.dat';

%let catlength=5;      /*For fixed-length CAT - set number of items to administer for each dim*/

%let corr=0.0;

%let sigmatrix = {1  0.0  0.0,
                  0.0  1  0.0,
                  0.0  0.0  1}; /*for three dimensions*/

*%let sigmatrix = {1}; /*for one dimension*/

%let est=2;           /* Ability estimation Method
                        1=ML (Newton-Raphson) 2=MAP (Newton-Raphson) */

%let itemsel=2;       /* item selection Method
                        1=maxinfo (Newton-Raphson) 2=Baysian (Newton-Raphson) */

%let d=1.0;

%let ndims=3;         /* Define the number of trait dimensions */

%let modelType=2; /* 1 for 3P model; 2 for graded model*/

%let nparms=7;        /* Number of parameters across all items */

%let macrofile='Codes_final\Multi_Macros.sas'; * Name of Macro file with program modules;

filename wrkdir 'c:\MIRT';

%include wrkdir(&macrofile);      /* Load macros */

/*put the logout to file of NEWOUT, used when too many examinees*/

```

```
FILENAME NEWOUT 'c:\MIRT\output\temp.log';
```

```
PROC PRINTTO log=NEWOUT;
```

```
run;
```

```
%itempar;
```

```
/*use either one for item response*/
```

```
%Response_gen;
```

```
*%Response_get;
```

```
%catloop;
```

```
%output;
```

```
quit;
```

```
*****,
```

```
* MACRO FILES WHICH ARE USED IN MAIN FILE
```

```
*****,
```

```
%macro itempar;
```

```
%global nitems;
```

```
/* Step 1 of 3: Extract item parameter information */
```

```
data item_par_full;
```

```
infile wrkdir(&parmfile) missover;
```

```
input x1-x%eval(&nparms) dim ncat;
```

```
call symput('npool',_n_); /* Determine size of item pool */
```

```
run;
```

```
%let nitems=%eval(&npool);
```

```
data item_par;
```

```
set item_par_full;
```

```
array y{*} x1-x%eval(&nparms);
```

```

        keep p;
        do j=1 to %eval(&nparms);
            p=y{j};
            output;
        end;
run;

data category;
    set item_par_full;
    keep ncat;
run;

/* Create a row vector with item parameters as elements */
proc transpose data=item_par out=item_par prefix=p;
    var p;
run;

proc transpose data=category out=category prefix=ncat;
    var ncat;
run;

%mend itempar;

%macro Response_gen;
/* create multivariate normal theta */
proc iml;
    mu=repeat(0,&ndims,1);
    sigma=&sigmatrix;
    p=nrow(sigma);
    x=normal(repeat(0,&nExaminee,p));
    u=(root(sigma));
    b=repeat(mu`,&nExaminee,1);

```



```

    theta=(x*u)+b;
    create theta_true from theta;
    append from theta;
    close theta_true;
quit;

data theta_true;
    set theta_true;
    rename col1-col&ndims=true1-true&ndims;
run;

/* create uniform probability for each examinee and each item */
data probability;
    array u{*} u1-u&nitems;
    do i=1 to &nExaminee;
        do j=1 to &nitems;
            u[j] = UNIFORM(0);
        end;
        output;
    end;
    drop i j;
run;

/* create item response set */
data response;
    if _n_=1 then do; * Estimates and parameters can now be compared;
        set item_par;
        set category;
    end;
    set theta_true;
    set probability;

```

```

array true{*} true1-true&ndims;
array p{&nitems,%eval(&nparms)} p1-p%eval(&nitems * &nparms);
array resp{*} resp1-resp&nitems;
array pstar{*} pstar1-pstar10;
array u{*} u1-u&nitems;
array ncat{*} ncat1-ncat&nitems; /*number of categories*/
array pGr{*} pGr1-pGr10;

pstar[1] = 1;
do j=1 to &nitems;
    pstar[ncat[j]+1] = 0;
    if &modelType = 1 then do;
        %di_res;
        resp[j] = prob>u[j];
    end;
    else do;
        %gr_res;
        do i=2 to ncat[j]+1;
            pGr[i-1] = pstar[i-1]-pstar[i];
        end;
        do i=2 to ncat[j];
            pGr[i] = pGr[i-1]+pGr[i];
        end;
        resp[j]=0;
        do i=1 to ncat[j]-1;
            if u[j]>pGr[i] and u[j]<=pGr[i+1] then
                resp[j]=i;
        end;
    end;
end;
end;

```

```

        file wrkdir(&responsefile);
        put (resp1-resp&nitems) (1.0);
        keep resp1-resp&nitems;
run;
%mend Response_gen;

%macro Response_get;
data response;
    infile wrkdir(&responsefile);
    input (resp1-resp30) (1.0) true1 true2 true3;
run;

data theta_true;
    set response;
    keep true1-true3;
run;

data response;
    set response;
    keep resp1-resp30;
run;
%mend Response_get;

%macro di_res;
    bparm = p{j,%eval(&ndims+1)}; /*b*/
    argmnt = 0;
    do tht = 1 to &ndims;
        argmnt = argmnt + p{j,tht}*(true(tht)-bparm);
    end;
    cparm = p{j,%eval(&ndims+2)};
    prob = cparm + (1 - cparm)/(1 + exp(-&d*( argmnt )));

```

```

%mend di_res;

%macro gr_res;
  do i=1 to ncat[j]-1;
    argmnt = 0;
    do tht=1 to &ndims;
      argmnt = argmnt + p[j,tht]*(true[tht]-p[j,&ndims+i]);
    end;
    pstar[i+1] = 1/(1+exp(-&d*argmnt));
  end;
%mend;

%macro di_cat;
  bparm = p[admin[c,j],%eval(&ndims+1)];
  argmnt = 0;
  do i = 1 to &ndims;
    argmnt = argmnt + p[admin[c,j],i]*(t[i]-bparm);
  end;
  cparm = p[admin[c,j],%eval(&ndims+2)];
  prob[j] = cparm + (1 - cparm)/(1 + exp(-&d*( argmnt )));
%mend di_cat;

%macro gr_cat;
  do i=1 to ncat[admin[c,j]]-1;
    argmnt = 0;
    do tht=1 to &ndims;
      argmnt = argmnt + p[admin[c,j],tht]*(t[tht]-p[admin[c,j],&ndims+i]);
    end;
    pstar[i+1] = 1/(1+exp(-&d*argmnt));
  end;
%mend gr_cat;

```

```

%macro di_sel;
  bparm = p[j,%eval(&ndims+1)];
  argmnt = 0;
  do i = 1 to &ndims;
    argmnt = argmnt + p[j,i]*(theta[c,i]-bparm);
  end;
  cparm = p[j,%eval(&ndims+2)];

  prob[j] = cparm + (1 - cparm)/(1 + exp(-&d*( argmnt )));
%mend di_sel;

%macro gr_sel;
  do i=1 to ncat[j]-1;
    argmnt = 0;
    do tht=1 to &ndims;
      argmnt = argmnt + p[j,tht]*(theta[c,tht]-p[j,&ndims+i]);
    end;
    pstar[i+1] = 1/(1+exp(-&d*argmnt));
  end;
%mend gr_sel;

%macro di_prov;
  bparm = p[admin[c,j],%eval(&ndims+1)];
  argmnt = 0;
  do i = 1 to &ndims;
    argmnt = argmnt + p[admin[c,j],i]*(theta[c,i]-bparm);
  end;
  cparm = p[admin[c,j],%eval(&ndims+2)];

  prob[j] = cparm + (1 - cparm)/(1 + exp(-&d*( argmnt )));

```

```
%mend di_prov;
```

```
%macro gr_prov;
```

```
do i=1 to ncat[admin[c,j]]-1;
```

```
    argmnt = 0;
```

```
    do tht=1 to &ndims;
```

```
        argmnt = argmnt + p[admin[c,j],tht]*(theta[c,tht]-p[admin[c,j],&ndims+i]);
```

```
    end;
```

```
pstar[i+1] = 1/(1+exp(-&d*argmnt));
```

```
end;
```

```
%mend gr_prov;
```

```
%macro catloop;
```

```
proc iml;
```

```
admin = j(%eval(&nExaminee),%eval(&catlength),0.);
```

```
theta = j(%eval(&nExaminee),%eval(&ndims),0.);
```

```
provse = j(%eval(&nExaminee),%eval(&ndims),0.);
```

```
total = j(%eval(&nExaminee),1,0.);
```

```
resp = j(%eval(&nExaminee),%eval(&catlength),0.);
```

```
dim = j(%eval(&nExaminee),%eval(&ndims),0.);
```

```
pstar = j(1,10,0.);
```

```
pstar[1] = 1;
```

```
use response;
```

```
read all var _all_ into ix;
```

```
use item_par_full;
```

```
read all var _all_ into p;
```

```
use category;
```

```
read all var _all_ into ncat;
```

```

start f_likelihood(t) global(resp,p,icounter,admin,c,ncat);
prob = j(1,%eval(&nitems),0.);
sum = 0.;

do j = 1 to icounter;
    pstar = j(1,10,0.);
    pstar[1] = 1;
    if &modelType = 1 then do;
        %di_cat;
        sum = sum + resp[c,j]*log(prob[j]) + (1-resp[c,j])*log(1-prob[j]);
    end;
    else do;
        %gr_cat;
        k=resp[c,j]+2;
        lnL = pstar[k-1]-pstar[k];
        if lnL = 0 then lnL = 0.000001;
        sum = sum + log(lnL);
    end;
end;

if &est = 2 then do;
    mu=repeat(0,&ndims,1);
    sigma1=&sigmatrix;
    sigma=Inv(sigma1);
    multi = j(1,%eval(&ndims),0.);

    do j=1 to &ndims;
        do i=1 to &ndims;
            multi[j] = multi[j] + (t[i]-mu[i])*sigma[i,j];
        end;
    end;
end;

```

```

sum1 = 0.;
do j=1 to &ndims;
    sum1 = sum1 + multi[j]*(t[j]-mu[j]);
end;
sum = sum - sum1/2 - &ndims/2*log(2*3.14) - 1/2*log(det(signal));
end;

f = sum;
return(f);

finish f_likelihood;

start g_likelihood(t) global(resp,p,icounter,admin,c,ncat);
    prob = j(1,%eval(&nitems),0.);
    g = j(1,%eval(&ndims),0.);

    do j = 1 to icounter;
        pstar = j(1,10,0.);
        pstar[1] = 1;
        if &modelType = 1 then do;
            %di_cat;
            do i = 1 to &ndims;
                g[i] = g[i]&d*p[admin[c,j],i]*(prob[j]-cparm)*(resp[c,j]-prob[j])/((1-
cparm)*prob[j]);
            end;
        end;
        else do;
            %gr_cat;
            k=resp[c,j]+2;
            if pstar[k-1]-pstar[k] = 0 then pstar[k-1]-pstar[k] = 0.000001;

```



```

        dlnLdt = pstar[k-1]*(1-pstar[k-1])-pstar[k]*(1-pstar[k]);
        dlnLdt = dlnLdt/(pstar[k-1]-pstar[k]);
        do i = 1 to &ndims;
            g[i] = g[i]+&d*p[admin[c,j],i]*dlnLdt;
        end;
    end;
end;

if &est = 2 then do;
    mu=repeat(0,&ndims,1);
    sigma=&sigmatrix;
    sigma=Inv(sigma);

    do j=1 to &ndims;
        sum = 0.;
        do i=1 to &ndims;
            sum = sum + sigma[j,i]*(t[i]-mu[i]);
        end;
        g[j] = g[j] - sum;
    end;
end;

return(g);

finish g_likelihood;

start h_likelihood(t) global(resp,p,icounter,admin,c,ncat);
    prob = j(1,%eval(&nitems),0.);
    h=j(%eval(&ndims),%eval(&ndims),0.);

    do j = 1 to icounter;

```

```

pstar = j(1,10,0.);
pstar[1] = 1;
if &modelType = 1 then do;
    %di_cat;
    prob[j]=(&d**2)*(1-prob[j])*(prob[j]-cparm)*(cparm*resp[c,j]-
prob[j]**2)/((prob[j]*(1-cparm))**2);
end;
else do;
    %gr_cat;
    k=resp[c,j]+2;
    prob[j] = pstar[k-1]*(1-pstar[k-1])+pstar[k]*(1-pstar[k]);
    prob[j] = -&d**2*prob[j];
end;
end;

do j= 1 to icounter;
    do i=1 to &ndims;
        do k=1 to &ndims;
            h[i,k]=h[i,k]+p[admin[c,j],i]*p[admin[c,j],k]*prob[admin[c,j]];
        end;
    end;
end;

if &est = 2 then do;
    sigma=&sigmatrix;
    sigma=Inv(sigma);
    h = h - sigma;
end;
return(h);

finish h_likelihood;

```

```

con1 = { -4,
        4 }; /*Parameter Constraints*/
do i=1 to &ndims;
    con = con || con1;
end;

do icounter=1 to &catlength;
    %check;    /*for check use*/
    %itemselect;
    %abilityEst;

    if mod(icounter,5)=0 then do;
        %provSE;
        if icounter=5 then do;
            Est = theta;
            SE = provse;
            SETotal = total;
        end;
        else do;
            Est = Est||theta;
            SE = SE||provse;
            SETotal = SETotal||total;
        end;
    end;
end;

create estimate from Est;
append from Est;
close estimate;

```

```

create SE from SE;
append from SE;
close SE;

create SETotal from SETotal;
append from SETotal;
close SETotal;

create admin from admin;
append from admin;
close admin;

quit;
%mend catloop;

%macro check;
do c = 1 to &nExaminee;
    admin[c,icounter] = icounter;
    resp[c,icounter] = ix[c,icounter];
end;
%mend;

%macro itemselect;
do c = 1 to &nExaminee;

    prob = j(1,%eval(&nitems),0.);

    do j = 1 to &nitems;
        pstar = j(1,10,0.);
        pstar[1] = 1;
        if &modelType = 1 then do;

```

```

        %di_sel;
        prob[j]=(&d**2)*(1-prob[j])*((prob[j]-cparm)**2)/(prob[j]*(1-
cparm)**2);
    end;
    else do;
        %gr_sel;
        do i=2 to ncat[j]+1;
            prob[j]=prob[j]+(pstar[i-1]*(1-pstar[i-1])+pstar[i]*(1-
pstar[i]))*(pstar[i-1]-pstar[i]);
        end;
        prob[j]=(&d**2)*prob[j];
    end;
end;

sigma=&sigmatrix;
sigma=Inv(sigma);

infoMatrix=j(%eval(&ndims),%eval(&ndims),0.0);
info = j(1,%eval(&nitems),0.);

I_thetaj = j(%eval(&ndims),%eval(&ndims),0.);

if &ndims > 1 then do;
    do j= 1 to icounter-1;
        do i=1 to &ndims;
            do k=1 to &ndims;

I_thetaj[i,k]=I_thetaj[i,k]+p[admin[c,j],i]*p[admin[c,j],k]*prob[admin[c,j]];
            end;
        end;
    end;
end;

```

```

end;

do i=1 to icounter-1;
    info[admin[c,i]] = -1;
end;

if icounter>0 & icounter<6 then do; start=1; myEnd=10; end;
else if icounter>5 & icounter<11 then do; start=11; myEnd=20; end;
else if icounter>10 & icounter<16 then do; start=21; myEnd=30; end;

do j=start to myEnd;
    if info[j] = 0 then do;
        do i=1 to &ndims;
            do k=1 to &ndims;
                infoMatrix[i,k]=p[j,i]*p[j,k]*prob[j];
            end;
        end;
        if &itemsel=1 then info[j]=det(I_thetaj+infoMatrix);
        else info[j]=det(I_thetaj+infoMatrix+sigma);
    end;
end;

maxvalue=0;
selected=1;

do search=start to myEnd;
    if info[search]>maxvalue then do;
        maxvalue=info[search];
        selected=search;
    end;
end;

```

```

        admin[c,icounter] = selected;
        resp[c,icounter] = ix[c,selected];

end;
%mend itemselect;

%macro abilityEst;

    theta0 = j(1,&ndims,0.);
    optn = {1 0}; /*1 means maximazition; 0 means no print*/
    tc = {. . . 0.01};

    do c = 1 to &nExaminee;
        call
nlpnra(rc,thetares,"f_likelihood",theta0,optn,con,tc,,,"g_likelihood","h_likelihood");
        if c=1 then t = thetares;
        else t = t//thetares;
        fopt = f_likelihood(t);
    end;

    theta = t;

%mend abilityEst;

%macro provSE;

    do c = 1 to &nExaminee;
        I_theta = j(%eval(&ndims),%eval(&ndims),0.);
        prob = j(1,%eval(&nitems),0.);

        do j = 1 to icounter;

```

```

        if &modelType = 1 then do;
            %di_prov;
            prob[j]=(&d**2)*(1-prob[j])*((prob[j]-
cparm)**2)/(prob[j]*(1-cparm)**2);
        end;
        else do;
            %gr_prov;
            do i=2 to ncat[admin[c,j]];
                prob[j]=prob[j]+(pstar[i-1]-pstar[i])*(1-
pstar[i-1]-pstar[i])**2;
            end;
            prob[j]=(&d**2)*prob[j];
        end;
        do i=1 to &ndims;
            do k=1 to &ndims;

I_theta[i,k]=I_theta[i,k]+p[admin[c,j],i]*p[admin[c,j],k]*prob[j];
            end;
        end;
        end;
        sigma=&sigmatrix;
        sigma=Inv(sigma);
        total[c] = det(I_theta + sigma);

        do i=1 to &ndims;
            provse[c,i]=I_theta[i,i];
        end;
    end;

%mend;
%macro output;

```


Data estimate;

set estimate;

rename col1-col%eval(&catlength/5) = est1-est%eval(&catlength/5);

run;

Data SE;

set SE;

rename col1-col%eval(&catlength/5) = SE1-SE%eval(&catlength/5);

run;

Data SETotal;

set SETotal;

rename col1-col%eval(&catlength/15) = SETotal1-SETotal%eval(&catlength/15);

run;

Data admin;

set admin;

file wrkdir(&adminfile);

put (col1-col%eval(&catlength)) (4.0);

run;

Data output;

merge theta_true estimate SE SETotal;

array est{*} est1-est%eval(&catlength/5);

array true{*} true1-true&ndims;

array SE{*} SE1-SE%eval(&catlength/5);

file wrkdir(&outputfile);

put (true1-true%eval(&ndims) est1-est%eval(&ndims)) (7.3);

run;

%mend;

BIBLIOGRAPHY

- Ackerman T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied psychological measurement*, 20, 311-330.
- Ackerman T. A., & Gierl M. J., & Walker C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. NCME instructional module.
- Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66, 541-562.
- Bjorner, J.B., & Ware, J.E. (1998). Using modern psychometric methods to measure health outcomes. *Medical Outcomes Trust Monitor*, 3(2), 11-16.
- Bold, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, 27, 395-414.
- Bock, R. D, Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.
- Chen, S.-Y., Ankenmann, R.D., & Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, 24, 241-255.
- Embretson, S.E., & Reise, S.P. Item response theory for psychologists. Mahwah: New Jersey: Lawrence Erlbaum.
- Fraser, C. (1988). *NOHARM*. [Computer software and manual]. Armidale, New South Wales, Australia: author.
- Haley, S.M., Mchorney, C.A., & Ware, J.E. (1994). Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *Journal of Clinical Epidemiology*, 47, 671-684.
- Hambleton, R.K., & Swaminathan, H. (1985). Item response theory. Boston, MA.
- Harwell, M, Stone, C.A., Hsu, T.C., & Kirisci, L. (1996) Monte Carlo Studies in Item Response Theory. *Applied Psychological Measurement*, 20, 101-126.

- Hojtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, 62, 171-189.
- Irrgang, J.J., Stone, C.A. (2004). *Differential item function for the disabilities of the arm shoulder and hand (DASH) outcome measure*. Presented at the Combined Sections Meeting of the American Physical Therapy Association, Nashville TN.
- Irrgang, J.J., Stone, C.A. (in press). Differential item function for the disabilities of the arm shoulder and hand (DASH) outcome measure. *Journal of Orthopaedic and Sports Physical Therapy*.
- Kingsbury, G.G. & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Psychological Measurement*, 4, 241-261.
- Lee, H. (1995). Markov Chain Monte Carlo methods for estimating multidimensional ability in item response analysis (Doctoral dissertation, University of Missouri, Columbia, MO).
- Li H. Y. & Schafer W. D. (2004). The context effects of multidimensional CAT on the accuracy of multidimensional abilities. *Paper presented at AERA 2004*.
- Luecht R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied psychological measurement*, 20, 389-404.
- McKinley, R. & Rechase, M.D. (1983). An extension of the two-parameter logistic model to the multidimensional latent space. (Research Report ONR83-2). Iowa City, IA: American College Testing Program.
- McHorney, C.A. (1997). Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. *Annals of Internal Medicine*, 127, 743-750.
- McHorney, C.A. & Cohen, A.S. (2000). Equating health status measures with item response theory: Illustration with functional status items. *Medical Care*, 38 (9 Supplement), 43-59.
- Muraki, E. & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, 19, 73-90.
- Muraki E. & Engelhard Jr. G. (1985). Full-information item factor analysis: applications of EAP scores. *Applied psychological measurement*, 9, 417-430.
- Ostini, R. & Nering M. L. *Polytomous item response theory models*, Thousand Oaks, CA: Sage Publications.
- Parshall C.G. & Spray J.A. & Kalohn J.C. & Davey T. (2002). *Practical Considerations in Computer-Based Testing*. Springer: New York.

- Reckase, M.D. (1997). A linear logistic multidimensional item response model for dichotomous item response data. In Van der Linder, W.J. & Hambleton, R.K. (eds). *Handbook of Modern Item Response Theory*. New York: Springer.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, 17.
- Samejima, F. (1972). A general model for free response data. *Psychometric Monograph*, 18.
- Segall, D.O. (1996). Multidimensional adaptive testing. *The psychometric society*, 61, 331-354.
- Segall, D.O. (2000). Principles of multidimensional adaptive testing. In van der Linder W.J. & Glass C.A.W. (eds.). *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers.
- Solway, S., Beaton, D.E., McConnell, S., & Bombardier C. (2002). The DASH Outcome Measure user's manual. Toronto, Ontario, Canada: The Institute for work & Health. Rosemont, Illinois: The American Academy of Orthopaedic Surgeons.
- Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-292.
- Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools. (Report No. ETS-RR-94-5). Princeton, NJ: ETS.
- Stone, C. A. (1993). *The use of multiple replications in IRT based Monte Carlo research*. Paper presented at European Meeting of the Psychometric Society, Barcelon.
- Stone, C.A., & Irrgang, J.J. (2004). Simulation of a computer adaptive test utilizing the disabilities of the arm, shoulder and hand (DASH) outcome measure. Presented at the Combined Sections Meeting of the American Physical Therapy Association, Nashville TN.
- Stone, C.A., & Weissman, A. (2002). Simulating computer adaptive tests. Final Report for University of Pittsburgh, Central Research Development Fund (August 2002).
- Sympson, J. B. (1978). A model for testing multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference*. 82-98.
- Torre, J.D.L. & Patz, R.J. (2002). A multidimensional item response theory approach to simultaneous ability estimation. Presented in NCME.
- Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0). Mooresville, IN: Scientific Software.
- Urry, V. W. (1977). Tailor testing: A successful application of item response theory. *Journal of Educational Measurement*, 14, 181-196.

- Van der Linden, W. J. (1997). Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Wainer, H. (1990). *Computer adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.
- Waller, N. G., & Reise, S.P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology*, 57, 1050-1058.
- Wang, T. & Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 109-135.
- Ware, J. E., Bjorner, J.B., & Kosinsky, M. (2000). Practical implications of item response theory and computerized adaptive testing. In D.J. Lubinski, & R. V. Davis (Eds), *Assessing individual differences in human behavior: New concepts, methods, and findings* (49-79). Palo Alto, CA: Davies-Black Publishing.
- Ware, J. E., Kosinski, M., & Keller, D.K. (1994). *SF-36 physical and mental health summary scales: A user's manual*. Boston, MA: The Health Institute.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika*, 54, 427-450.
- Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17, 17-27.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.
- Weiss, D. J. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* 49-79, Palo Alto CA: Davies-Black Publishing.
- Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479-494.